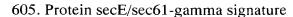605. Protein secE/sec61-gamma signature

In bacteria, the secE protein plays a role in protein export; it is one of the components - with

secY and secA - of the preprotein translocase. In eukaryotes, the evolutionary related protein

5   sec61-gamma playsa role in protein translocation through the endoplasmic reticulum; it is

part of a trimeric complex that also consist of sec61-alpha and beta [1]. Both secE and sec61-

gamma are small proteins of about 60 to 90 amino acids that contain a single transmembrane

region at their C-terminal extremity (Escherichia colisecE is an exception, in that it possess

an extra N-terminal segment of 60residues that contains two additional transmembrane

10   domains).The sequence of secE/sec61-gamma is not extremely well conserved, however it is

possible to derive a signature pattern centered on a conserved proline located 10 residues

before the beginning of the transmembrane domain.

Consensus pattern: [LIVMFY]-x(2)-[DENQGA]-x(4)-[LIVMFTA]-x-[KRV]-x(2)-[KW]-P-
x(3)-[SEQ]-x(7)-[LIVT]-[LIVGA]-[LIVFGAST]

15   [1] Hartmann E., Sommer T., Prehn S., Goerlich D., Jentsch S., Rapoport T.A. Nature
367:654-657(1994).

606. 11-S plant seed storage proteins signature

20   Plant seed storage proteins, whose principal function appears to be the major nitrogen source

for the developing plant, can be classified, on the basis of their structure, into different

families. 11-S are non-glycosylated proteins which form hexameric structures [1,2]. Each of

the subunits in the hexamer is itself composed of an acidic and a basic chain derived from a

single precursor and linked by a disulfide bond. This structure is shown in the following

25   representation. +-----------------------+ | |

xxxxxxxxxxxCxxxxxxxxxxxxxxxxxxxxxxxxNGxCxxxxxxxxxxxxxxxxxxxxxxxx ********* <--

----Acidic-subunit-------------><-----Basic-subunit------> <----------------About-480-to-500-

residues----------------->'C': conserved cysteine involved in a disulfide bond.'*': position of the

pattern. Proteins that belong to the 11-S family are: pea and broad bean legumins, rape

30   cruciferin, rice glutelins, cotton beta-globulins, soybean glycinins, pumpkin 11-S globulin,

oat globulin, sunflower helianthinin G3, etc. The region that includes the conserved cleavage

site between the acidic and basic subunits (Asn-Gly) and a proximal cysteine residue which is

involved in the interchain disulfide bond have been used as a signature pattern for this family

of proteins.

Consensus pattern: N-G-x-[DE](2)-x-[LIVMF]-C-[ST]-x(11,12)-[PAG]-D [C is involved in a disulfide bond

[1] Hayashi M., Mori H., Nishimura M., Akazawa T., Hara-Nishimura I. Eur. J. Biochem. 172:627-632(1988).[2] Shotwell M.A., Afonso C., Davies E., Chesnut R.S., Larkins B.A. Plant Physiol. 87:698-704(1988).

607. 7S seed storage protein

7S globulin is one of the main storage proteins of most angiosperms and gymnosperms. The 7S storage proteins are homotrimers.

Number of members: 67

[1] The three-dimensional structure of canavalin from jack bean (Canavalia ensiformis). Ko TP, Ng JD, McPherson A; Plant Physiol 1993;101:729-744.

608. Aspartate-semialdehyde dehydrogenase signature

Aspartate-semialdehyde dehydrogenase (ASD) catalyzes the second step in the common biosynthetic pathway leading from Asp to diaminopimelate and Lys, to Met, and to Thr; the NADP-dependent reductive dephosphorylation of L-aspartyl phosphate to L-aspartate-semialdehyde. In bacteria and fungi, ASDis a protein of about 40 Kd (340 to 370 residues) whose sequence is not extremely well conserved [1]. A conserved cysteine residue has been implicated as important for the catalytic activity [2].The region of conservation around the active site residue is too small to be used as signature pattern. Another more conserved region, located in the last third of the sequence, and which contains both a conserved cysteine as well as an histidine has been used instead.

Consensus pattern: [LIVM]-[SADN]-x(2)-C-x-R-[LIVM]-x(4)-[GSC]-H-[STA

[1] Baril C., Richaud C., Fourni E., Baranton G., Saint Girons I. J. Gen. Microbiol. 138:47-53(1992).[2] Karsten W.E., Viola R.E. Biochim. Biophys. Acta 1121:234-238(1992).

N-acetyl-gamma-glutamyl-phosphate reductase active site

N-acetyl-gamma-glutamyl-phosphate reductase (EC 1.2.1.38) (AGPR) [1,2] is the enzyme that catalyzes the third step in the biosynthesis of arginine from glutamate, the NADP-dependent reduction of N-acetyl-5-glutamyl phosphate into N-acetylglutamate 5-semialdehyde.In bacteria it is a monofunctional protein of 35 to 38 Kd (gene argC) while in

fungi it is part of a bifunctional mitochondrial enzyme (gene ARG5,6, arg11 orarg-6) which contains a N-terminal acetylglutamate kinase (EC 2.7.2.8) domain and a C-terminal AGPR domain. In the Escherichia coli enzyme, a cysteine has been shown to be implicated in the catalytic activity, the region around this residue is well conserved and can be used as a

5   signature pattern.

Consensus pattern: [LIVM]-[GSA]-x-P-G-C-[FY]-[AVP]-T-[GA]-x(3)-[GTAC]-[LIVM]- x- P [C is the active site residue]

[1] Ludovice M., Martin J.F., Carrachas P., Liras P. J. Bacteriol. 174:4606-4613(1992).[2] Gessert S.F., Kim J.H., Nargang F.E., Weiss R.L. J. Biol. Chem. 269:8189-8203(1994).

10

609. Sialyltransferase family,

Number of members: 18

15

610. SpoU rRNA Methylase family

This family of proteins probably use S-AdoMet. Number of members: 58

[1] SpoU protein of Escherichia coli belongs to a new family of putative rRNA methylases. Koonin EV, Rudd KE; Nucleic Acids Res 1993;21:5519-5519. [2] The spoU gene of

20   escherichia coli , the fourth gene of the spoT operon, is essential for tRNA (Gm18) 2 ' methyltransferase activity. Persson BC, Jager G, Gustafsson C; Nucleic Acids Res 1997;25:4093-4097.

25   611. Stathmin family signatures

Stathmin [1] (from the Greek 'stathmos' which means relay), is an ubiquitous intracellular protein, present in a variety of phosphorylated forms and which serves as a relay for diverse second messenger pathways. Its expression and phosphorylation are regulated throughout development and in response to extracellular signals regulating cell proliferation,

30   differentiation and function. Stathmin is a highly conserved protein of 149 amino acid residues. Structurally, it consists of an N-terminal domain of about 45 residues followed by a 78 residue alpha-helical domain consisting of a heptad repeat coiled coil structure and a C-terminal domain of 25 residues. Protein SCG10 is a neuron-specific, membrane-associated protein that accumulates in the growth cones of developing neurons. It is highly similar in its

sequence to stathmin, but differs in that it contains an additional N-terminal hydrophobic segment of 32 residues which is probably responsible for its interaction with membranes. Xenopus protein XB3 is also evolutionary related to stathmin and also contains an additional N-terminal hydrophobic domain [2]. A conserved decapeptide which ends with the first three residues of the coiled coil domain and a second pattern that corresponds to part of the central region of the coiled coil have been selected as signatures for proteins of the stathmin family.

Consensus pattern: P-[KRQ]-[KR](2)-[DE]-x-S-L-[EG]-E-

Consensus pattern: A-E-K-R-E-H-E-[KR]-E-

[1] Sobel A. Trends Biochem. Sci. 16:301-305(1991).[2] Maucuer A., Moreau J., Mechali M., Sobel A. J. Biol. Chem. 268:16420-16429(1993).

612. SUA5/yciO/yrdC family signature. The following uncharacterized proteins have been shown [1] to share regions of similarities: - Yeast protein SUA5. - Escherichia coli hypothetical protein yciO and HI1198, the corresponding Haemophilus influenzae protein. - Escherichia coli hypothetical protein yrdC and HI0656, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein ywlC. - Mycobacterium leprae hypothetical protein in rfe-hemK intergenic region. - Methanococcus jannaschii hypothetical protein MJ0062.These are proteins of from 20 to 46 Kd which contain a number of conserved regions in their N-terminal section. They can be picked up in the database by the following pattern.

Consensus pattern: [LIVMTA](3)-[LIVMFYC]-[PG]-T-[DE]-[STA]-x-[FY]-[GA]- [LIVM]-[GS]-

[1] Bairoch A., Rudd K.E., Robison K. Unpublished observations (1995).

613. Sucrose synthase

Sucrose synthases catalyse the synthesis of sucrose from UDP-glucose and fructose. This family includes the bulk of the sucrose synthase protein. However the carboxyl terminal region of the sucrose synthases belongs to the glycosyl transferase family Glycos_transf_1.
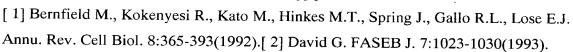
614. Sulfotransferase proteins

Number of members: 59

615. Synaptophysin / synaptoporin signature

Synaptophysin and synaptoporin [1] are structurally related proteins, found in the membrane of synaptic vesicles, which may function as ionic or solute channels. These two glycoproteins seem to span the membrane four times. Both their N- and C-termini sequences seem to be cytoplasmically located. As a signature pattern for this family of proteins, a highly conserved region located in the beginning of the first intravesicular loop just after the first transmembrane domain has been selected. This region contains a cysteine residue that may be involved in a disulfide bond.

Consensus pattern: L-S-V-[DE]-C-x-N-K-T [C may be involved in a disulfide bond

[1] Knaus P., Marqueze-Pouey B., Scherer H., Betz H. Neuron 5:453-462(1990).

616. Syndecans signature

Syndecans [1,2] (from the greek syndein; to bind together) are a family of transmembrane heparan sulfate proteoglycans which are implicated in the binding of extracellular matrix components and growth factors. Syndecans bind a variety of molecules via their heparan sulfate chains and can act as receptors or as co-receptors. Structurally, these proteins consist of four separate domains: a) A signal sequence; b) An extracellular domain (ectodomain) of variable length and whose sequence is not evolutionary conserved in the various forms of syndecans. The ectodomain contains the sites of attachment of the heparan sulfate glycosaminoglycan side chains; c) A transmembrane region; d) A highly conserved cytoplasmic domain of about 30 to 35 residues which could interact with cytoskeletal proteins. The proteins known to belong to this family are: - Syndecan 1. - Syndecan 2 or fibroglycan. - Syndecan 3 or neuroglycan or N-syndecan. - Syndecan 4 or amphiglycan or ryudocan. - Drosophila syndecan. - Caenorhabditis elegans probable syndecan (F57C7.3).The signature pattern that has been developed for syndecans starts with the last residue of the transmembrane region and includes the first 10 residues of the cytoplasmic domain. This region, which contains four basic residues, could act as a stop transfer site.

Consensus pattern: [FY]-R-[IM]-[KR]-K(2)-D-E-G-S-Y

[1] Bernfield M., Kokenyesi R., Kato M., Hinkes M.T., Spring J., Gallo R.L., Lose E.J. Annu. Rev. Cell Biol. 8:365-393(1992).[2] David G. FASEB J. 7:1023-1030(1993).

5   617. Syntaxin / epimorphin family signature

The following proteins have been shown to be evolutionary related [1,2,3]: - Epimorphin (or syntaxin 2), a mammalian mesenchymal protein which plays an essential role in epithelial morphogenesis. - Syntaxin 1A (also known as antigen HPC-1) and syntaxin 1B which are synaptic proteins which may be involved in docking of synaptic vesicles at presynaptic active

10   zones. - Syntaxin 3. - Syntaxin 4, which is potentially involved in docking of synaptic vesicles at presynaptic active zones. - Syntaxin 5, which mediates endoplasmic reticulum to golgi transport. - Syntaxin 6, which is involved in intracellular vesicle trafficking. - Syntaxin 7. - Yeast PEP12 (or VPS6) which is required for the transport of proteases to the vacuole. - Yeast SED5 which is required for the fusion of transport vesicles with the Golgi complex. -

15   Yeast SSO1 and SSO2 which are required for vesicle fusion with the plasma membrane. - Yeast VAM3, which is required for vacuolar assembly. - Arabidopsis thaliana protein KNOLLE which may be involved in cytokinesis. - Caenorhabditis elegans hypothetical proteins F35C8.4, F48F7.2, F55A11.2 and T01B11.3.The above proteins share the following characteristics: a size ranging from30 Kd to 40 Kd; a C-terminal extremity which is highly

20   hydrophobic and isprobably involved in anchoring the protein to the membrane; a central, well conserved region, which seems to be in a coiled-coil conformation. The pattern specific for this family is based on the most conserved region of the coiled coil domain.

Consensus pattern: [RQ]-x(3)-[LIVMA]-x(2)-[LIVM]-[ESH]-x(2)-[LIVMT]-x-[DEVM]-[LIVM]-x(2)-[LIVM]-[FS]-x(2)-[LIVM]-x(3)-[LIVT]-x(2)-Q- [GADEQ]-x(2)-[LIVM]-

25   [DNQT]-x-[LIVMF]-[DESV]-x(2)-[LIVM]

[1] Bennett M.K., Garcia-Arraras J.E., Elferink L.A., Peterson K., Fleming A.M., Hazuka C.D., Scheller R.H. Cell 74:863-873(1993).[2] Spring J., Kato M., Bernfield M. Trends Biochem. Sci. 18:124-125(1993).[3] Pelham H.R.B. Cell 73:425-426(1993).

30

618. Sm protein

The U1, U2, U4/U6, and U5 small nuclear ribonucleoprotein particles (snRNPs) involved in pre-mRNA splicing contain seven Sm proteins (B/B', D1, D2, D3, E, F and G) in common, which

assemble around the Sm site present in four of the major spliceosomal small nuclear RNAs. These proteins contain a common sequence motif in two segments, Sm1 and Sm2, separated by a short variable linker.

5

[1] Hermann H, Fabrizio P, Raker VA, Foulaki K, Hornig H, Brahms H, Luhrmann R EMBO J 1995;14:2076-2088. [2] Kambach C, Walke S, Young R, Avis JM, de la Fortelle E, Raker VA, Luhrmann R, Li J, Nagai K; Cell 1999;96:375-387.

10

619. Skp1 family

[1] Stebbins CE, Kaelin WG Jr, Pavletich NP; Science 1999;284:455-461.

15

620. Protein secY signatures

The eubacterial secY protein [1] plays an important role in protein export. It interacts with the signal sequences of secretory proteins as well as with two other components of the protein translocation system: secA and secE. SecY is an integral plasma membrane protein of 419 to

20   492 amino acid residues that apparently contains ten transmembrane segments. Such a structure probablyconfers to secY a 'translocator' function, providing a channel for periplasmic and outer-membrane precursor proteins.Homologs of secY are found in archaebacteria [2]. SecY is also encoded in the chloroplast genome of some algae [3] where it could be involved in a prokaryotic-like protein export system across the two membranes of

25   the chloroplast endoplasmic reticulum (CER) which is present in chromophyte andcryptophyte algae. Two signature patterns have been developed for secY proteins. The first corresponds to the second transmembrane region, which is the most conserved section of these proteins. The second spans the C-terminal part of the fourth transmembrane region, a short intracellular loop, and the N-terminal part of the fifth transmembrane region.

30   Consensus pattern: [GST]-[LIVMF](2)-x-[LIVM]-G-[LIVM]-x-P-[LIVMFY](2)-x-[AS]-[GSTQ]-[LIVMFAT](3)-Q-[LIVMFA](2)

Consensus pattern: [LIVMFYW](2)-x-[DE]-x-[LIVMF]-[STN]-x(2)-G-[LIVMF]-[GST]-[NST]-G-x-[GST]-[LIVMF](3)

[1] Ito K. Mol. Microbiol. 6:2423-2428(1992).[2] Auer J., Spicker G., Boeck A. Biochimie 73:683-688(1991).[3] Douglas S.E. FEBS Lett. 298:93-96(1992).

5   621. (Seed protein) Small hydrophilic plant seed proteins signature. The following small hydrophilic plant seed proteins are structurally related: - Arabidopsis thaliana proteins GEA1 and GEA6. - Cotton late embryogenesis abundant (LEA) protein D-19. - Carrot EMB-1 protein. - Barley LEA proteins B19.1A, B19.1B, B19.3 and B19.4. - Maize late embryogenesis abundant protein Emb564. - Radish late seed maturation protein p8B6. - Rice

10   embryonic abundant protein Emp1. - Sunflower 10 Kd late embryogenesis abundant protein (DS10). - Wheat Em proteins. These proteins contains from 83 to 153 amino acid residues and may play a role[1,2] in equipping the seed for survival, maintaining a minimal level of hydration in the dry organism and preventing the denaturation of cytoplasmic components. They may also play a role during imbibition by controlling water uptake. As a signature

15   pattern, the best conserved region in the sequence of these proteins has been developed, it is a glycine-rich nonapeptide located in the N-terminal section.-

Consensus pattern: G-[EQ]-T-V-V-P-G-G-T-

20   [1] Dure L. III, Crouch M., Harada J., Ho T.-H. D., Mundy J., Quatrano R., Thomas T., Sung Z.R. Plant Mol. Biol. 12:475-486(1989).[2] Gaubier P., Raynal M., Hull G., Huestis G.M., Grellet F., Arenas C., Pages M., Delseny M. Mol. Gen. Genet. 238:409-418(1993).

25   622. Serine carboxypeptidases, active sites
All known carboxypeptidases are either metallo carboxypeptidases or serinecarboxypeptidases. The catalytic activity of the serine carboxypeptidases, like that of the trypsin family serine proteases, is provided by a charge relay system involving an aspartic acid residue hydrogen-bonded to a histidine, which is itself hydrogen-bonded to a serine [1].

30   Proteins known to be serine carboxypeptidases are: - Barley and wheat serine carboxypeptidases I, II, and III [2]. - Yeast carboxypeptidase Y (YSCY) (gene PRC1), a vacuolar protease involved in degrading small peptides. - Yeast KEX1 protease, involved in killer toxin and alpha-factor precursor processing. - Fission yeast sxa2, a probable carboxypeptidase involved in degrading or processing mating pheromones [3]. - Penicillium

janthinellum carboxypeptidase S1 [4]. - Aspergullus niger carboxypeptidase pepF. - Aspergullus satoi carboxypeptidase cpdS. - Vertebrate protective protein / cathepsin A [5], a lysosomal protein which is not only a carboxypeptidase but also essential for the activity of both beta-galactosidase and neuraminidase. - Mosquito vitellogenic carboxypeptidase (VCP)

5   [6]. - Naegleria fowleri virulence-related protein Nf314 [7]. - Yeast hypothetical protein YBR139w. - Caenorhabditis elegans hypothetical proteins C08H9.1, F13D12.6, F32A5.3, F41C3.5 and K10B2.2.This family also includes: - Sorghum (s)-hydroxymandelonitrile lyase (hydroxynitrile lyase) (HNL) [8], an enzyme involved in plant cyanogenesis. The sequences surrounding the active site serine and histidine residues are highly conserved in all these

10   serine carboxypeptidases.

Consensus pattern: [LIVM]-x-[GTA]-E-S-Y-[AG]-[GS] [S is the active site residue]
Consensus pattern: [LIVF]-x(2)-[LIVSTA]-x-[IVPST]-x-[GSDNQL]-[SAGV]-[SG]-H-x-[IVAQ]-P-x(3)-[PSA] [H is the active site residue]

[1] Liao D.I., Remington S.J. J. Biol. Chem. 265:6528-6531(1990).[2] Sorensen S.B.,

15   Svendsen I., Breddam K. Carlsberg Res. Commun. 54:193-202(1989).[3] Imai Y., Yamamoto M. Mol. Cell. Biol. 12:1827-1834(1992).[4] Svendsen I., Hofmann T., Endrizzi J., Remington J., Breddam K. FEBS Lett. 333:39-43(1993).[5] Galjart N.J., Morreau H., Willemsen R., Gillemans N., Bonten E.J., d'Azzo A. J. Biol. Chem. 266:14754-14762(1991).[6] Cho W.L., Deitsch K.W., Raikhel A.S. Proc. Natl. Acad. Sci. U.S.A. 88:10821-

20   10824(1991).[7] Hu W.N., Kopachik W., Band R.N. Infect. Immun. 60:2418-2424(1992).[8] Wajant H., Mundry K.W., Pfitzenmaier K. Plant Mol. Biol. 26:735-746(1994).[9] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).[E1]

25   623. Serpins signature. Serpins (SERine Proteinase INhibitors) [1,2,3,4] are a group of structurally related proteins. They are high molecular weight (400 to 500 amino acids),extracellular, irreversible serine protease inhibitors with a well defined structural-functional characteristic: a reactive region that acts as a 'bait' for an appropriate serine protease. This region is found in the C-terminal part of these proteins. Proteins which are

30   known to belong to the serpin family are listed below (references are only provided for recently determined sequences): - Alpha-1 protease inhibitor (alpha-1-antitrypsin, contrapsin). - Alpha-1-antichymotrypsin, - Antithrombin III. - Alpha-2-antiplasmin. - Heparin cofactor II. - Complement C1 inhibitor. - Plasminogen activator inhibitors 1 (PAI-1) and 2 (PAI-2). - Glia derived nexin (GDN) (Protease nexin I). - Protein C inhibitor. - Rat

hepatocytes SPI-1, SPI-2 and SPI-3 inhibitors. - Human squamous cell carcinoma antigen (SCCA) which may act in the modulation of the host immune response against tumor cells. - A lepidopteran protease inhibitor. - Leukocyte elastase inhibitor which, in contrast to other serpins, is an intracellular protein. - Neuroserpin [5], a neuronal inhibitor of plasminogen

5   activators and plasmin. - Cowpox virus crmA [6], an inhibitor of the thiol protease interleukin-1B converting enzyme (ICE). CrmA is the only serpin known to inhibit a non-serine proteinase. - Some orthopoxviruses probable protease inhibitors, which may be involved in the regulation of the blood clotting cascade and/or of the complement cascade in the mammalian host. On the basis of strong sequence similarities, a number of proteins with

10   no known inhibitory activity are said to belong to this family: - Birds ovalbumin and the related genes X and Y proteins. - Angiotensinogen; the precursor of the angiotensin active peptide. - Barley protein Z; the major endosperm albumin. - Corticosteroid binding globulin (CBG). - Thyroxine-binding globulin (TBG). - Sheep uterine milk protein (UTMP) and pig uteroferrin-associated protein (UFAP). - Hsp47, an endoplasmic reticulum heat-shock protein

15   that binds strongly to collagen and could act as a chaperone in the collagen biosynthetic pathway [7]. - Maspin, which seems to function as a tumor supressor [5]. - Pigment epithelium-derived factor precursor (PEDF), a protein with a strong neutrophic activity [8]. - Ep45, an estrogen-regulated protein from Xenopus [9]. A signature pattern has been developed for this family of proteins, centered on a well conserved Pro-Phe sequence which

20   is found ten to fifteen residues on the C-terminal side of the reactive bond

Consensus pattern: [LIVMFY]-x-[LIVMFYAC]-[DNQ]-[RKHQS]-[PST]-F-[LIVMFY]-[LIVMFYC]-x-[LIVMFAH]-

25   [1] Carrell R., Travis J. Trends Biochem. Sci. 10:20-24(1985).[2] Carrell R., Pemberton P.A., Boswell D.R. Cold Spring Harbor Symp. Quant. Biol. 52:527-535(1987).[3] Huber R., Carrell R.W. Biochemistry 28:8951-8966(1989).[4] Remold-O'Donneel E. FEBS Lett. 315:105-108(1993).[5] Osterwalder T., Contartese J., Stoeckli E.T., Kuhn T.B., Sonderegger P. EMBO J. 15:2944-2953(1996).[6] Komiyama T., Ray C.A., Pickup D.J., Howard A.D.,

30   Thornberry N.A., Peterson E.P., Salvesen G. J. Biol. Chem. 269:19331-19337(1994).[7] Clarke E., Sandwal B.D. Biochim. Biophys. Acta 1129:246-248(1992).[8] Zou Z., Anisowicz A., Neveu M., Rafidi K., Sheng S., Sager R., Hendrix M.J., Seftor E., Thor A. Science 263:526-529(1994).[9] Steele F.R., Chader G.J., Johnson L.V., Tombran-Tink J.

5 624. Sigma-54 interaction domain signatures and profile

Some bacterial regulatory proteins activate the expression of genes from promoters recognized by core RNA polymerase associated with the alternative sigma-54 factor. These have a conserved domain of about 230 residues involved in the ATP-dependent [1,2] interaction with sigma-54. This domain has been found in the proteins listed below: - acoR

10 from Alcaligenes eutrophus, an activator of the acetoin catabolism operon acoXABC. - algB from Pseudomonas aeruginosa, an activator of alginate biosynthetic gene algD. - dctD from Rhizobium, an activator of dctA, the C4-dicarboxylate transport protein. - dhaR from Citrobacter freundii, a regulator of the dha operon for glycerol utilization. - fhlA from Escherichia coli, an activator of the formate dehydrogenase H and hydrogenase III structural

15 genes. - flbD from Caulobacter crescentus, an activator of flagellar genes. - hoxA from Alcaligenes eutrophus, an activator of the hydrogenase operon. - hrpS from Pseudomonas syringae, an activator of hprD as well as other hrp loci involved in plant pathogenicity. - hupR1 from Rhodobacter capsulatus, an activator of the [NiFe] hydrogenase genes hupSL. - hydG from Escherichia coli and Salmonella typhimurium, an activator of the hydrogenase

20 activity. - levR from Bacillus subtilis, which regulates the expression of the levanase operon (levDEFG and sacC). - nifA (as well as anfA and vnfA) from various bacteria, an activator of the nif nitrogen-fixing operon. - ntrC, from various bacteria, an activator of nitrogen assimilatory genes such as that for glutamine synthetase (glnA) or of the nif operon. - pgtA from Salmonella typhimurium, the activator of the inducible phospho- glycerate transport

25 system. - pilR from Pseudomonas aeruginosa, an activator of pilin gene transcription. - rocR from Bacillus subtilis, an activator of genes for arginine utilization - tyrR from Escherichia coli, involved in the transcriptional regulation of aromatic amino-acid biosynthesis and transport. - wtsA, from Erwinia stewartii, an activator of plant pathogenicity gene wtsB. - xylR from Pseudomonas putida, the activator of the tol plasmid xylene catabolism operon

30 xylCAB and of xylS. - Escherichia coli hypothetical protein yfhA. - Escherichia coli hypothetical protein yhgB. About half of these proteins (algB, dcdT, flbD, hoxA, hupR1, hydG, ntrC, pgtA and pilR) belong to signal transduction two-component systems [3] and possess a domain that can be phosphorylated by a sensor-kinase protein in their N- terminal section. Almost all of these proteins possess a helix-turn-helix DNA-binding domain in their

C-terminal section. The domain which interacts with the sigma-54 factor has an ATPase

activity. This may be required to promote a conformational change necessary for

theinteraction [4]. The domain contains an atypical ATP-binding motif A (P-loop) as well as

a form of motif B. The two ATP-binding motifs are located in the N-terminal section of the

5   domain; signature patterns have been developed for both motifs. Other regions of the domain

are also conserved. One of them, located in the C-terminal section, has been selected as a

third signature pattern.

Consensus pattern: [LIVMFY](3)-x-G-[DEQ]-[STE]-G-[STAV]-G-K-x(2)-[LIVMFY]

Consensus pattern: [GS]-x-[LIVMF]-x(2)-A-[DNEQASH]-[GNEK]-G-[STIM]-

10   [LIVMFY](3)-[DE]-[EK]-[LIVM]

Consensus pattern: [FYW]-P-[GS]-N-[LIVM]-R-[EQ]-L-x-[NHAT]

[1] Morrett E., Segovia L. J. Bacteriol. 175:6067-6074(1993).[2] Austin S., Kundrot C.,

Dixon R. Nucleic Acids Res. 19:2281-2287(1991).[3] Albright L.M., Huala E., Ausubel

F.M. Annu. Rev. Genet. 23:311-336(1989).[4] Austin S., Dixon R. EMBO J. 11:2219-

15   2228(1992).

625. Sigma-70 factors family signatures

Sigma factors [1] are bacterial transcription initiation factors that promote the attachment of

20   the core RNA polymerase to specific initiation sites and arethen released. They alter the

specificity of promoter recognition. Most bacteria express a multiplicity of sigma factors.

Two of these factors, sigma-70 (gene rpoD), generally known as the major or primary sigma

factor, and sigma-54 (gene rpoN or ntrA) direct the transcription of a wide variety of genes.

The other sigma factors, known as alternative sigma factors, are required for the transcription

25   of specific subsets of genes. With regard to sequence similarity, sigma factors can be grouped

into two classes: the sigma-54 and sigma-70 families. The sigma-70 family includes, in

addition to the primary sigma factor, a wide variety of sigma factors, some of which are listed

below: - Bacillus sigma factors involved in the control of sporulation-specific genes: sigma-E

(sigE or spoIIGB), sigma-F (sigF or spoIIAC), sigma-G (sigG or spoIIIG), sigma-H (sigH or

30   spo0C) and sigma-K (sigK or spoIVCB/spoIIIC). - Escherichia coli and related bacteria

sigma-32 (gene rpoH or htpR) involved in the expression of heat shock genes. - Escherichia

coli and related bacteria sigma-27 (gene fliA) involved in the expression of the flagellin gene.

- Escherichia coli sigma-S (gene rpoS or katF) which seems to be involved in the expression

of genes required for protection against external stresses. - Myxococcus xanthus sigma-B

(sigB) which is essential for the late-stage differentiation of that bacteria. Alignments of the sigma-70 family permit the identification of four regions of high conservation [2,3]. Each of these four regions can in turn be subdivided into a number of sub-regions. Signature patterns based on the two best-conserved sub-regions have been developed. The first pattern

5 corresponds to sub-region 2.2;the exact function of this sub-region is not known although it could be involved in the binding of the sigma factor to the core RNA polymerase. The second pattern corresponds to sub-region 4.2 which seems to harbor a DNA-binding 'helix-turn-helix' motif involved in binding the conserved -35region of promoters recognized by the major sigma factors. The second pattern starts one residue before the N-terminal extremity of the

10 HTH region and ends six residues after its C-terminal extremity.

Consensus pattern: [DE]-[LIVMF](2)-[HEQS]-x-G-x-[LIVMFA]-G-L-[LIVMFYE]-x-[GSAM]-[LIVMAP]

Consensus pattern: [STN]-x(2)-[DEQ]-[LIVM]-[GAS]-x(4)-[LIVMF]-[PSTG]-x(3)-[LIVMA]-x-[NQR]-[LIVMA]-[EQH]-x(3)-[LIVMFW]-x(2)-[LIVM]

15 [1] Helmann J.D., Chamberlin M.J. Annu. Rev. Biochem. 57:839-872(1988).[2] Gribskov M., Burgess R.R. Nucleic Acids Res. 14:6745-6763(1986).[3] Lonetto M.A., Gribskov M., Gross C.A. J. Bacteriol. 174:3843-3849(1992).[4] Lonetto M.A., Brown K.L., Rudd K.E., Buttner M.J. Proc. Natl. Acad. Sci. U.S.A. 91:7573-7577(1994).

20

626. Signal carboxyl-terminal domain. 430 members.

627. Signal peptidases I signatures

25 Signal peptidases (SPases) [1] (also known as leader peptidases) remove the signal peptides from secretory proteins. In prokaryotes three types of Spases are known: type I (gene lepB) which is responsible for the processing of the majority of exported pre-proteins; type II (gene lsp) which only process lipoproteins, and a third type involved in the processing of pili subunits. SPase I is an integral membrane protein that is anchored in the cytoplasmic

30 membrane by one (in B. subtilis) or two (in E. coli) N-terminal transmembrane domains with the main part of the protein protuding in the periplasmic space. Two residues have been shown [2,3] to be essential for the catalytic activity of SPase I: a serine and an lysine.SPase I is evolutionary related to the yeast mitochondrial inner membrane protease subunit 1 and 2 (genes IMP1 and IMP2) which catalyze the removal of signal peptides required for the

targeting of proteins from the mitochondrial matrix, across the inner membrane, into the inter-membrane space [4]. In eukaryotes the removal of signal peptides is effected by an oligomeric enzymatic complex composed of at least five subunits: the signal peptidase complex (SPC). The SPC is located in the endoplasmic reticulum membrane. Two

5   components of mammalian SPC, the 18 Kd (SPC18) and the 21 Kd (SPC21) subunits as well as the yeast SEC11 subunit have been shown [5] to share regions of sequence similarity with prokaryotic SPases I and yeast IMP1/IMP2. Three signature patterns for these proteins have been developed. The first signature contains the putative active site serine, the second signature contains the putative active site lysine which is not conserved in the SPC subunits,

10 and the third signature corresponds to a conserved region of unknown iological significance which is located in the C-terminal section of all these proteins.

Consensus pattern: [GS]-x-S-M-x-[PS]-[AT]-[LF] [S is an active site residue]

Consensus pattern: K-R-[LIVMSTA](2)-G-x-[PG]-G-[DE]-x-[LIVM]-x-[LIVMFY] [K is an active site residue]

15 Consensus pattern: [LIVMFYW](2)-x(2)-G-D-[NH]-x(3)-[SND]-x(2)-[SG]

[1] Dalbey R.E., von Heijne G. Trends Biochem. Sci. 17:474-478(1992).[2] Sung M., Dalbey R.E. J. Biol. Chem. 267:13154-13159(1992).[3] Black M.T. J. Bacteriol. 175:4957-4961(1993).[4] Nunnari J., Fox T.D., Walter P. Science 262:1997-2004(1993).[5] van Dijl J.M., de Jong A., Vehmaanpera J., Venema G., Bron S. EMBO J. 11:2819-2828(1992).[6]

20 Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).[E1]

628. (sodcu) Copper/Zinc superoxide dismutase signatures

Copper/Zinc superoxide dismutase (SODC) [1] is one of the three forms of an enzyme that

25 catalyzes the dismutation of superoxide radicals. SODC binds one atom each of zinc and copper. Various forms of SODC are known: acytoplasmic form in eukaryotes, an additional chloroplast form in plants, an extracellular form in some eukaryotes, and a periplasmic form in prokaryotes. The metal binding sites are conserved in all the known SODC sequences [2]. Two signature patterns have been derived for this family of enzymes: the first one contains

30 two histidine residues that bind the copper atom; the second one islocated in the C-terminal section of SODC and contains a cysteine which is involved in a disulfide bond.

Consensus pattern: [GA]-[IMFAT]-H-[LIVF]-H-x(2)-[GP]-[SDG]-x-[STAGDE] [The two H's are copper ligands]

Consensus pattern: G-[GN]-[SGA]-G-x-R-x-[SGA]-C-x(2)-[IV] [C is involved in a disulfide bond]

[1] Bannister J.V., Bannister W.H., Rotilio G. CRC Crit. Rev. Biochem. 22:111-154(1987).[2] Smith M.W., Doolittle R.F. J. Mol. Evol. 34:175-184(1992).

5

629. (sodfe) Manganese and iron superoxide dismutases signature

Manganese superoxide dismutase (SODM) [1] is one of the three forms of an enzyme that catalyzes the dismutation of superoxide radicals. The four ligands of the manganese atom are
10   conserved in all the known SODM sequences. These metal ligands are also conserved in the related iron form ofsuperoxide dismutases [2,3]. A short conserved region which includes two of the four ligands: an aspartate and a histidine has been selected as a signature.

Consensus pattern: D-x-W-E-H-[STA]-[FY](2) [D and H are manganese/iron ligands]

[1] Bannister J.V., Bannister W.H., Rotilio G. CRC Crit. Rev. Biochem. 22:111-154(1987).[
15   2] Parker M.W., Blake C.C.F. FEBS Lett. 229:377-382(1988).[3] Smith M.W., Doolittle R.F. J. Mol. Evol. 34:175-184(1992).

630. Spectrin repeat
20   Spectrin repeats are found in several proteins involved in cytoskeletal structure. These include spectrin, alpha-actinin and dystrophin.The sequence repeat used in this family is taken from the structural repeat in reference [2]. The spectrin repeat forms a three helix bundle. The second helix is interrupted by proline in some sequences.

25   Number of members: 898

[1] Actin-binding proteins. 1: Spectrin super family. Hartwig JH; Protein Profile 1995;2:732-732. [2] Crystal structure of the repetitive segments of spectrin. Yan Y, Winograd E, Viel A, Cronin T, Harrison SC, Branton D; Science 1993;262:2027-2030.

30

631. (subtilase) Streptomyces subtilisin-type inhibitors signature
Bacteria of the Streptomyces family produce a family of proteinase inhibitors[1] characterized by their strong activity toward subtilisin. They arecollectively known as SSI's: Streptomyces Subtilisin Inhibitors. Some SSI'salso inhibit trypsin or chymotrypsin. In their

mature secreted form, SSI's areproteins of about 110 residues with two conserved disulfide

bonds. +-------+ +--------------+ | | | |

xxxxxxxxxxxxxxCxxxxxxxCxxxxxxxxxCx#xxxxxxxxxxxxCxxxxxx ************'C':

conserved cysteine involved in a disulfide bond.'#': active site residue.'*': position of the

5   pattern.

Consensus pattern: C-x-P-x(2,3)-G-x-H-P-x(4)-A-C-[ATD]-x-L [The two C's are involved in

a disulfide bond]

[1] Taguchi S., Kojima S., Terabe M., Miura K.-I., Momose H. Eur. J. Biochem. 220:911-

918(1994).

10

632. Sugar transport proteins signatures

In mammalian cells the uptake of glucose is mediated by a family of closely related transport

proteins which are called the glucose transporters [1,2,3].At least seven of these transporters

15   are currently known to exist (in Human they are encoded by the GLUT1 to GLUT7

genes).These integral membrane proteins are predicted to comprise twelve membrane

spanning domains. The glucose transporters show sequence similarities [4,5] with a number

of other sugar or metabolite transport proteins listed below (references are only provided for

recently determined sequences). - Escherichia coli arabinose-proton symport (araE). -

20   Escherichia coli galactose-proton symport (galP). - Escherichia coli and Klebsiella

pneumoniae citrate-proton symport (also known as citrate utilization determinant) (gene cit).

- Escherichia coli alpha-ketoglutarate permease (gene kgtP). - Escherichia coli

proline/betaine transporter (gene proP) [6]. - Escherichia coli xylose-proton symport (xylE). -

Zymomonas mobilis glucose facilitated diffusion protein (gene glf). - Yeast high and low

25   affinity glucose transport proteins (genes SNF3, HXT1 to HXT14). - Yeast galactose

transporter (gene GAL2). - Yeast maltose permeases (genes MAL3T and MAL6T). - Yeast

myo-inositol transporters (genes ITR1 and ITR2). - Yeast carboxylic acid transporter protein

homolog JEN1. - Yeast inorganic phosphate transporter (gene PHO84). - Kluyveromyces

lactis lactose permease (gene LAC12). - Neurospora crassa quinate transporter (gene Qa-y),

30   and Emericella nidulans quinate permease (gene qutD). - Chlorella hexose carrier (gene

HUP1). - Arabidopsis thaliana glucose transporter (gene STP1). - Spinach sucrose

transporter. - Leishmania donovani transporters D1 and D2. - Leishmania enriettii probable

transport protein (LTP). - Yeast hypothetical proteins YBR241c, YCR98c and YFL040w. -

Caenorhabditis elegans hypothetical protein ZK637.1. - Escherichia coli hypothetical proteins

yabE, ydjE and yhjE. - Haemophilus influenzae hypothetical proteins HI0281 and HI0418. - Bacillus subtilis hypothetical proteins yxbC and yxdF. It has been suggested [4] that these transport proteins have evolved from theduplication of an ancestral protein with six transmembrane regions, this hypothesis is based on the conservation of two G-R-[KR] motifs.

5　The first one is located between the second and third transmembrane domains and the second one between transmembrane domains 8 and 9. Two patterns have been developed to detect this family of proteins. The first pattern is based on the G-R-[KR] motif; but because this motif is too short to be specific to this family of proteins, a pattern from a larger region centered on the second copy of this motif was derived. The second pattern is based on a

10　number of conserved residues which are located at the end of the fourth transmembrane segment and in the short loop region between the fourth and fifth segments.

Consensus pattern: [LIVMSTAG]-[LIVMFSAG]-x(2)-[LIVMSA]-[DE]-x-[LIVMFYWA]- G- R-[RK]-x(4,6)-[GSTA]

Consensus pattern: [LIVMF]-x-G-[LIVMFA]-x(2)-G-x(8)-[LIFY]-x(2)-[EQ]-x(6)- [RK]

15　[1] Silverman M. Annu. Rev. Biochem. 60:757-794(1991).[2] Gould G.W., Bell G.I. Trends Biochem. Sci. 15:18-23(1990).[3] Baldwin S.A. Biochim. Biophys. Acta 1154:17-49(1993).[4] Maiden M.C.J., Davis E.O., Baldwin S.A., Moore D.C.M., Henderson P.J.F. Nature 325:641-643(1987).[5] Henderson P.J.F. Curr. Opin. Struct. Biol. 1:590-601(1991).[6] Culham D.E., Lasby B., Marangoni A.G., Milner J.L., Steer B.A., van Nues R.W., Wood

20　J.M. J. Mol. Biol. 229:268-276(1993).

633. Synaptobrevin signature

Synaptobrevin [1] is an intrinsic membrane protein of small synaptic vesicles whose function

25　is not yet known, but which is highly conserved in mammals, electric ray (where its is known as VAMP-1), Drosophila and yeast [2]. In yeast there are two closely related forms of synaptobrevin (genes SNC1 andSNC2) while in mammals there is at least 4 (genes SYB1, SYB2, SYB3 and SYBL1).Structurally synaptobrevin consist of a N-terminal cytoplasmic domain of from 90 to 110 residues, followed by a transmembrane region, and then by a short

30　(from 2 to 22 residues) C-terminal intravesicular domain. As a signature pattern for synaptobrevin, a highly conserved stretch of residues located in the central part of the sequence was selected.

Consensus pattern: N-[LIVM]-[DENS]-[KL]-V-x-[DEQ]-R-x(2)-[KR]-[LIVM]-[STDE]- x- [LIVM]-x-[DE]-[KR]-[TA]-[DE]

[1] Suedhof T.C., Baumert M., Perin M.S., Jahn R. Neuron 2:1475-1481(1989).[2] Gerst J.E., Rodgers L., Riggs M., Wigler M. Proc. Natl. Acad. Sci. U.S.A. 89:4338-4342(1992).

634. TBC domain. Identification of a TBC domain in GYP6_YEAST and GYP7_YEAST, which are GTPase activator proteins of yeast Ypt6 and Ypt7, imply that these domains are GTPase activator proteins of Rab-like small GTPases. Number of members: 55

[1] Medline: 96032578. Molecular cloning of a cDNA with a novel domain present in the tre-2 oncogene and the yeast cell cycle regulators BUB2 and cdc16. Richardson PM, Zon LI; Oncogene 1995;11:1139-1148.

[2]Medline: 97398935. A shared domain between a spindle assembly checkpoint protein and Ypt/Rab-specific GTPase-activators. Neuwald AF; Trends Biochem Sci 1997;22:243-244.

635. Transcription factor TFIID repeat signature (TBP)

Transcription factor TFIID (or TATA-binding protein, TBP) [1,2] is a general factor that plays a major role in the activation of eukaryotic genes transcribed by RNA polymerase II. TFIID binds specifically to the TATA box promoter element which lies close to the position of transcription initiation. There is a remarkable degree of sequence conservation of a C-terminal domain of about 180 residues in TFIID from various eukaryotic sources. This region isnecessary and sufficient for TATA box binding. The most significant structural feature of this domain is the presence of two conserved repeats of a 77 amino-acid region. The intramolecular symmetry generates a saddle-shaped structure that sits astride the DNA [3]. Drosophila TRF (TBP-related factor) [4] is a sequence-specific transcription factor that also binds to the TATA box and is highly similar to TFIID. Archaebacteria also possess a TBP homolog [5]. A signature pattern that spans the last 50 residues of the repeated region has been derived.-

Consensus pattern: Y-x-P-x(2)-[IF]-x(2)-[LIVM](2)-x-[KRH]-x(3)-P-[RKQ]-x(3)- L-[LIVM]-F-x-[STN]-G-[KR]-[LIVM]-x(3)-G-[TAGL]-[KR]-x(7)- [AGC]-x(7)-[LIVM

[1] Hoffmann A., Sinn E., Yamamoto T., Wang J., Roy A., Horikoshi M., Roeder R.G. Nature 346:387-390(1990).[2] Gash A., Hoffmann A., Horikoshi M., Roeder R.G., Chua N.-H. Nature 346:390-394(1990).[3] Nikolov D.B., Hu S.-H., Lin J., Gasch A., Hoffmann A., Horikoshi M., Chua N.-H., Roeder R.G., Burley S.K. Nature 360:40-46(1992).[4] Crowley T.E., Hoey T., Liu J.-K., Jan Y.N., Jan L.Y., Tjian R. Nature 361:557-561(1993).[5] Marsh

T.L., Reich C.I., Whitelock R.B., Olsen G.J. Proc. Natl. Acad. Sci. U.S.A. 91:4180-4184(1994).

5   636. Translationally controlled tumor protein signatures (TCTP)

Mammalian translationally controlled tumor protein (TCTP) (or P23) is a protein which has been found to be preferentially synthesized in cells during the early growth phase of some types of tumor [1,2], but which is also expressed in normal cells. The physiological function of TCTP is still not known. It is a hydrophilic protein of 18 to 20 Kd. Close homologs have

10   been found in plants [3], earthworm [4], Caenorhabditis elegans (F52H2.11), Hydra, budding yeast (YKL056c) [5] and fission yeast (SpAC1F12.02c) Two of the best conserved regions have been selected as signature patterns for TCTP.

Consensus pattern: [IFA]-[GA]-[GAS]-N-[PAK]-S-[GA]-E-[GDE]-[PAGE]-[DEQGA]

Consensus pattern: [FLVH]-[FY]-[IVCT]-G-E-x-[MA]-x(2,5)-[DEN]-[GAST]-x-[LV]-

15   [AV]-x(3)-[FYW]

[1] Boehm H., Beendorf R., Gaestel M., Gross B., Nuernberg P., Kraft R., Otto A., Bielka H. Biochem. Int. 19:277-286(1989).[2] Makrides S., Chitpatima S.T., Bandyopadhyay R., Brawerman G. Nucleic Acids Res. 16:2350-2350(1988).[3] Pay A., Heberle-Bors E., Hirt H. Plant Mol. Biol. 19:501-503(1992).[4] Stuerzenbaum S.R., Kille P., Morgan A.J. Biochim.

20   Biophys. Acta 1398:294-304(1998).[5] Rasmussen S.W. Yeast 10:S63-S68(1994).

637. TFIIS zinc ribbon domain signature

Transcription factor S-II (TFIIS) [1] is a eukaryotic protein necessary for efficient RNA

25   polymerase II transcription elongation, past template-encoded pause sites. TFIIS shows DNA-binding activity only in the presence of RNA polymerase II. It is a protein of about 300 amino acids whose sequence is highly conserved in mammals, Drosophila, yeast (where it was first known as PPR2, a transcriptional regulator of URA4, and then as DST1, the DNA strand transfer protein alpha [2]) and in the archaebacteria Sulfolobus acidocaldarius [3].This

30   family also includes the eukaryotic and archebacterial RNA polymerase subunits of the 15 Kd / M family (see <PDOC00790>) as well as the following viral proteins: - Vaccinia virus RNA polymerase 30 Kd subunit (rpo30) [4]. - African swine fever virus protein I243L [5].The best conserved region of all these proteins contains four cysteines that bind a zinc ion and fold in a conformation termed a 'zinc ribbon' [6]. Besides these cysteines, there are a

number of other conserved residues which can be used to help define a specific pattern for this type of domain.

Consensus pattern: C-x(2)-C-x(9)-[LIVMQSAR]-[QH]-[STQL]-[RA]-[SACR]-x-[DE]-[DET]-[PGSEA]-x(6)-C-x(2,5)-C-x(3)-[FW] [The four C's are zinc ligands]

5   [1] Hirashima S., Hirai H., Nakanishi Y., Natori S. J. Biol. Chem. 263:3858-3863(1988).[2] Kipling D., Kearsey S.E. Nature 353:509-509(1991).[3] Langer D., Zillig W. Nucleic Acids Res. 21:2251-2251(1993).[4] Ahn B.-Y., Gershon P.D., Jones E.V., Moss B. Mol. Cell. Biol. 10:5433-5441(1990).[5] Rodriguez J.M., Salas M.L., Vinuela E. Virology 186:40-52(1992).[6] Qian X., Jeon C., Yoon H., Agarwal K., Weiss M.A. Nature 365:277-279(1993).

10

638. Tetrahydrofolate dehydrogenase/cyclohydrolase signatures (THF DHG CYH)

Enzymes that participate in the transfer of one-carbon units are involved in various biosynthetic pathways. In many of these processes the transfers of one-carbon units are

15   mediated by the coenzyme tetrahydrofolate (THF). Various reactions generate one-carbon derivatives of THF which can be interconverted between different oxidation states by formyltetrahydrofolate synthetase(EC 6.3.4.3), methylenetetrahydrofolate dehydrogenase (EC 1.5.1.5 or EC 1.5.1.15) and methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9).The dehydrogenase and cyclohydrolase activities are expressed by a variety of multifunctional

20   enzymes: - Eukaryotic C-1-tetrahydrofolate synthase (C1-THF synthase), which catalyzes all three reactions described above. Two forms of C1-THF synthases are known [1], one is located in the mitochondrial matrix, while the second one is cytoplasmic. In both forms the dehydrogenase/cyclohydrolase domain is located in the N-terminal section of the 900 amino acids protein and consists of about 300 amino acid residues. The C1-THF synthases are

25   NADP- dependent. - Eukaryotic mitochondrial bifunctional dehydrogenase/cyclohydrolase [2]. This is an homodimeric NAD-dependent enzyme of about 300 amino acid residues. - Bacterial folD [3]. FolD is an homodimeric bifunctional NADP-dependent enzyme of about 290 amino acid residues. The sequence of the dehydrogenase/cyclohydrolase domain is highly conserved in all forms of the enzyme. Two conserved regions have been selected as

30   signature patterns. The first one is located in the N-terminal part of these enzymes and contains three acidic residues. The second pattern is a highly conserved sequence of 9 amino acids which is located in the C-terminal section.

Consensus pattern: [EQ]-x-[EQK]-[LIVM](2)-x(2)-[LIVM]-x(2)-[LIVMY]-N-x-[DN]- x(5)-[LIVMF](3)-Q-L-P-[LV]

Consensus pattern: P-G-G-V-G-P-[MF]-T-[IV]

[1] Shannon K.W., Rabinowitz J.C. J. Biol. Chem. 263:7717-7725(1988).[2] Belanger C., Mackenzie R.E. J. Biol. Chem. 264:4837-4843(1989).[3] d'Ari L., Rabinowitz J.C. J. Biol. Chem. 266:23953-23958(1991).

639. Triosephosphate isomerase active site (TIM)

Triosephosphate isomerase (EC 5.3.1.1) (TIM) [1] is the glycolytic enzyme that catalyzes the reversible interconversion of glyceraldehyde 3-phosphate and dihydroxyacetone phosphate. TIM plays an important role in several metabolic pathways and is essential for efficient energy production. It is a dimer of identical subunits, each of which is made up of about 250 amino-acid residues. A glutamic acid residue is involved in the catalytic mechanism [2]. The sequence around the active site residue is perfectly conserved in all known TIM's and can be used as a signature pattern for this type of enzyme.

Consensus pattern: [AV]-Y-E-P-[LIVM]-W-[SA]-I-G-T-[GK] [E is the active site residue]

[1] Lolis E., Alber T., Davenport R.C., Rose D., Hartman F.C., Petsko G.A. Biochemistry 29:6609-6618(1990).[2] Knowles J.R. Nature 350:121-124(1991).

640. Thymidine kinase cellular-type signature (TK)

Thymidine kinase (TK) (EC 2.7.1.21) is an ubiquitous enzyme that catalyzes the ATP-dependent phosphorylation of thymidine. A comparison of TK sequences has shown [1,2,3] that there are two different families of TK. One family groups together TK from herpes viruses as well as cellular thymidylate kinases, while the second family currently consists of TK from the following sources: - Vertebrates. - Bacterial. - Bacteriophage T4. – Pox viruses. - African swine fever virus (ASF). - Fish lymphocystis disease virus (FLDV).A conserved region which is located in the C-terminal section of these enzymes has been selected as a signature pattern for this family of TKA.

Consensus pattern: [GA]-x(1,2)-[DE]-x-Y-x-[STAP]-x-C-[NKR]-x-[CH]-[LIVMFYWH]

[1] Boyle D.B., Coupar B.E.H., Gibbs A.J., Seigman L.J., Both G.W. Virology 156:355-365(1987).[2] Blasco R., Lopez-Otin C., Munoz M., Bockamp E.-O., Simon-Mateo C., Vinuela E. Virology 178:301-304(1990).[3] Robertson G.R., Whalley J.M. Nucleic Acids Res. 16:11303-11317(1988).

641. Thymidine kinase from herpesvirus (TK herpes)

[1]

Medline: 96003730

5   Crystal structures of the thymidine kinase from herpes

simplex virus type-1 in complex with deoxythymidine and

ganciclovir.

   Brown DG, Visse R, Sandhu G, Davies A, Rizkallah PJ, Melitz

   C, Summers WC, Sanderson MR;

10   Nat Struct Biol 1995;2:876-881.

Number of members: 65

642. Nuclear transition protein 2 signatures (TP2)

15   In mammals, the second stage of spermatogenesis is characterized by the conversion of

nucleosomal chromatin to the compact, non-nucleosomal and transcriptionally inactive form

found in the sperm nucleus. This condensation is associated with a double-protein transition.

The first transition corresponds to the replacement of histones by several spermatid-specific

proteins, also called transition proteins, which are themselves replaced by protamines during

20   the second transition. Nuclear transition protein 2 (TP2) is one of those spermatid-specific

proteins. TP2 is a basic, zinc-binding protein [1] of 116 to 137 amino-acid residues.

Structurally, TP2 consists of three distinct parts: a conserved serine-rich N-terminal domain

of about 25 residues, a variable central domain of 20 to 50 residues which contains cysteine

residues, and a conserved C-terminal domain of about 70 residues rich in lysines and

25   arginines. Two signature patterns for TP2 have been developed: one located in the N-terminal

domain, the other in the C-terminal.

Consensus pattern: H-x(3)-H-S-[NS]-S-x-P-Q-S

Consensus pattern: K-x-R-K-x(2)-E-G-K-x(2)-K-[KR]-K

[1] Baskaran R., Rao M.R.S. Biochem. Biophys. Res. Commun. 179:1491-1499(1991).

30

643. Thiamine pyrophosphate enzymes signature (TTP enzymes)

A number of enzymes require thiamine pyrophosphate (TPP) (vitamin B1) as a cofactor. It

has been shown [1] that some of these enzymes are structurally related. These related TPP

enzymes are: - Pyruvate oxidase (POX) (EC 1.2.3.3) Reaction catalyzed: pyruvate +

orthophosphate + O(2) + H(2)O = acetyl phosphate + CO(2) + H(2)O(2). - Pyruvate

decarboxylase (PDC) (EC 4.1.1.1) Reaction catalyzed: pyruvate = acetaldehyde + CO(2). -

Indolepyruvate decarboxylase (EC 4.1.1.74) [2] Reaction catalyzed: indole-3-pyruvate =

5   indole-3-acetaldehyde + CO(2). - Acetolactate synthase (ALS) (EC 4.1.3.18) Reaction

catalyzed: 2 pyruvate = acetolactate + CO(2). - Benzoylformate decarboxylase (BFD) (EC

4.1.1.7) [3] Reaction catalyzed: benzoylformate = benzaldehyde + CO(2). A conserved

region which is located in their C-terminal section has been selected as a signature pattern for

these enzymes.

10   Consensus pattern: [LIVMF]-[GSA]-x(5)-P-x(4)-[LIVMFYW]-x-[LIVMF]-x-G-D-[GSA]-

[GSAC]

[1] Green J.B.A. FEBS Lett. 246:1-5(1989).[2] Koga J., Adachi T., Hidaka H. Mol. Gen.

Genet. 226:10-16(1991).[3] Tsou A.Y., Ransom S.C., Gerlt J.A., Buechter D.D., Babbitt

P.C., Kenyon G.L. Biochemistry 29:9856-9862(1990).

15

644. TPR Domain

     [1]

Medline: 95397415

20   Tetratrico peptide repeat interactions: to TPR or not to TPR?

     Lamb JR, Tugendreich S, Hieter P;

         Trends Biochem Sci 1995;20:257-259.

     [2]Medline: 98151343

     The structure of the tetratricopeptide repeats of protein

25   phosphatase 5: implications for TPR-mediated protein-protein

     interactions.

     Das AK, Cohen PW, Barford D;

         EMBO J 1998;17:1192-1199.

     Number of members: 621

30

645. Uroporphyrin-III C-methyltransferase signatures (TP methylase)

Uroporphyrin-III C-methyltransferase (EC 2.1.1.107) (SUMT) [1,2] catalyzes the transfer of

two methyl groups from S-adenosyl-L-methionine to the C-2 and C-7atoms of

uroporphyrinogen III to yield precorrin-2 via the intermediate formation of precorrin-1. SUMT is the first enzyme specific to the cobalamin pathway and precorrin-2 is a common intermediate in the biosynthesis of corrinoids such as vitamin B12, siroheme and coenzyme F430.The sequences of SUMT from a variety of eubacterial and archaebacterial species are

5 currently available. In species such as Bacillus megaterium (gene cobA), Pseudomonas denitrificans (cobA) or Methanobacterium ivanovii (gene corA) SUMT is a protein of about 25 to 30 Kd. In Escherichia coli and related bacteria, the cysG protein, which is involved in the biosynthesis of siroheme, is a multifunctional protein composed of a N-terminal domain, probably involved in transforming precorrin-2 into siroheme, and a C-terminal domain which

10 has SUMT activity. The sequence of SUMT is related to that of a number of P. denitrificans and Salmonella typhimurium enzymes involved in the biosynthesis of cobalamin which also seem to be SAM-dependent methyltransferases [3,4]. The similarity is especially strong with two of these enzymes: cobI/cbiL which encodes S-adenosyl-L-methionine--precorrin-2 methyltransferase and cobM/cbiF whose exact function is not known. Two signature patterns

15 have been developed for these enzymes. The first corresponds to a well conserved region in the N-terminal extremity (called region 1 in [1,3]) and the second to a less conserved region located in the central part of these proteins (this pattern spans what are called regions 2 and 3 in [1,3]).

Consensus pattern: [LIVM]-[GS]-[STAL]-G-P-G-x(3)-[LIVMFY]-[LIVM]-T-[LIVM]-

20 [KRHQG]-[AG]

Consensus pattern: V-x(2)-[LI]-x(2)-G-D-x(3)-[FYW]-[GS]-x(8)-[LIVF]-x(5,6)-[LIVMFYWPAC]-x-[LIVMY]-x-P-G

[1] Blanche F., Robin C., Couder M., Faucher D., Cauchois L., Cameron B., Crouzet J. J. Bacteriol. 173:4637-4645(1991).[2] Robin C., Blanche F., Cauchois L., Cameron B., Couder

25 M., Crouzet J. J. Bacteriol. 173:4893-4896(1991).[3] Crouzet J., Cameron B., Cauchois L., Rigault S., Rouyez M.-C., Blanche F., Thibaut D., Debussche L. J. Bacteriol. 172:5980-5990(1990).[4] Roth J.R., Lawrence J.G., Rubenfield M., Kieffer-Higgins S., Church G.M. J. Bacteriol. 175:3303-3316(1993).[5] Mattheakis L.C., Shen W.H., Collier R.J. Mol. Cell. Biol. 12:4026-4037(1992).

30

646. Tudor domain

Domain of unknown function present in several RNA-binding proteins. copies in the Drosophila Tudor protein. Slight ambiguities in the alignment.Number of members: 18

[1]Medline: 97200561 Tudor domains in proteins that interact with RNA. Ponting CP;
Trends Biochem Sci 1997;22:51-52. [2]Medline: 97157029 The human EBNA-2
coactivator p100: multidomain organization and relationship to the staphylococcal nuclease
fold and to the tudor protein involved in Drosophila melanogaster development. Callebaut I,
5 Mornon JP; Biochem J 1997;321:125-132.

647. Terpene synthase family

It has been suggested that this gene family be designated
10 tps (for terpene synthase) [1]. It has been split into six
subgroups on the basis of phylogeny, called tpsa-tpsf.
tpsa includes vetispiridiene synthase Swiss:Q39979, 5-epi-
aristolochene synthase, Swiss:Q40577 and (+)-delta-cadinene
synthase Swiss:P93665.
15 tpsb includes (-)-limonene synthase, Swiss:Q40322.
tpsc includes kaurene synthase A, Swiss:O04408.
tpsd includes taxadiene synthase, Swiss:Q41594, pinene synthase,
Swiss:O24475 and myrcene synthase, Swiss:O24474.
tpse includes kaurene synthase B.
20 tpsf includes linalool synthase.
Number of members: 51
[1]
Medline: 97413772
Monoterpene synthases from grand fir (Abies grandis). cDNA
25 isolation, characterization, and functional expression of
myrcene synthase, (-)-(4S)-limonene synthase, and
(-)-(1S,5S)-pinene synthase.
Bohlmann J, Steele CL, Croteau R;
J Biol Chem 1997;272:21784-21792.
30

648. ThiF family

This family contains a repeated domain in ubiquitin
activating enzyme E1 and members of the bacterial

ThiF/MoeB/HesA family.Number of members: 87

649. Thioester dehydrase

5   Members of this family are involved in fatty acid biosynthesis.

Number of members: 19

    [1]

Medline: 96398612

Structure of a dehydratase-isomerase from the bacterial

10   pathway for biosynthesis of unsaturated fatty acids: two

catalytic activities in one active site.

    Leesong M, Henderson BS, Gillig JR, Schwab JM, Smith JL;

       Structure 1996;4:253-264.

Database Reference: SCOP; 1mka; fa; [SCOP-USA][CATH-PDBSUM]

15   Database reference: PFAMB; PB058036;

650. Tub family signatures

The mouse tubby mutation is the cause of maturity-onset obesity, insulin resistance and

20   sensory deficits. This mutation maps to a gene, tub [1,2],which codes for a protein that

belongs to a family which currently consists of the following members: - Mammalian tub, an

hydrophilic protein of about 500 residues, which could be involved in the hypothalamic

regulation of body weight. - Human protein TULP1 [3] which may be involved in retinis

pigmentosa 14, a retinal degeneration disease. - Mouse protein p4-6 whose function is not

25   known. - Caenorhabditis elegans hypothetical protein F10B5.4. - Several fragmentary

sequences from plants, Drosophila and human ESTs. While the N-terminal part of these

protein is not conserved in length nor in the sequence, the C-terminal 250 residues are highly

conserved. Therefore, two regions were selected in the C-terminal part as signature patterns.

The secondr egion is located at the C-terminal extremity and contains a penultimate cysteine

30   residue that could be critical to the normal functioning of these proteins.

Consensus pattern: F-[KHQ]-G-R-V-[ST]-x-A-S-V-K-N-F-Q

Consensus pattern: A-F-[AG]-I-[SAC]-[LIVM]-[ST]-S-F-x-[GST]-K-x-A-C-E

[1] Kleyn P.W., Fan W., Kovats S.G., Lee J.L., Pulido J.C., Wu Y., Berkemeier L.R.,

Misumi D.J., Holmgren L., Charlat O., Woolf E.A., Tayber O., Brody T., Shu P., Hawkins F.,

Kennedy B., Baldini L., Ebeling C., Alperin G.D., Deeds J., Lakey N.D., Culpepper J., Chen H., Gluecksmann-Kuis M.A., Carlson G.A., Duyk G.M., Moore K.J. Cell 85:281-290(1996).[2] Noben-Trauth K., Naggert J.K., North M.A., Nishina P.M. Nature 380:534-538(1996).[3] North M.A., Naggert J.K., Yan Y., Noben-Trauth K., Nishina P.M. Proc. Natl. Acad. Sci. U.S.A. 94:3128-3133(1997).

651. Eukaryotic DNA topoisomerase I active site

DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type Itopoisomerases act by catalyzing the transient breakage of DNA, one strand at a time, and the subsequent rejoining of the strands. When a eukaryotic type Itopoisomerase breaks a DNA backbone bond, it simultaneously forms a protein-DNA link where the hydroxyl group of a tyrosine residue is joined to a 3'-phosphate on DNA, at one end of the enzyme-severed DNA strand. In eukaryotes and pox virus topoisomerases I, there are a number of conserved residues in the region around the active site tyrosine.

Consensus pattern: [DEN]-x(6)-[GS]-[IT]-S-K-x(2)-Y-[LIVM]-x(3)-[LIVM] [Y is the active site tyrosine]

[1] Sternglanz R. Curr. Opin. Cell Biol. 1:533-535(1990).[2] Sharma A., Mondragon A. Curr. Opin. Struct. Biol. 5:39-47(1995).[3] Lynn R.M., Bjornsti M.-A., Caron P.R., Wang J.C. Proc. Natl. Acad. Sci. U.S.A. 86:3559-3563(1989).[4] Roca J. Trends Biochem. Sci. 20:156-160(1995).[E1]

652. Transaldolase signatures

Transaldolase (EC 2.2.1.2) catalyzes the reversible transfer of a three-carbonketol unit from sedoheptulose 7-phosphate to glyceraldehyde 3-phosphate to form erythrose 4-phosphate and fructose 6-phosphate. This enzyme, together with transketolase, provides a link between the glycolytic and pentose-phosphate pathways. Transaldolase is an enzyme of about 34 Kd whose sequence has been well conserved throughout evolution. A lysine has been implicated [1]in the catalytic mechanism of the enzyme; it acts as a nucleophilic group that attacks the carbonyl group of fructose-6-phosphate.Transaldolase is evolutionary related [2] to a bacterial protein of about 20Kd (known as talC in Escherichia coli), whose exact function is not yet known. Two signature patterns have been developed for these proteins. The first,

located in the N-terminal section, contains a perfectly conserved pentapeptide; these cond, includes the active site lysine.

Consensus pattern: [DG]-[IVSA]-T-[ST]-N-P-[STA]-[LIVMF](2)

Consensus pattern: [LIVM]-x-[LIVM]-K-[LIVM]-[PAS]-x-[ST]-x-[DENQPAS]-G- [LIVM]-
5   x-[AGV]-x-[QEKRST]-x-[LIVM] [K is the active site residue]

[1] Miosga T., Schaaff-Gerstenschlaeger I., Franken E., Zimmermann F.K. Yeast 9:1241-
1249(1993).[2] Reizer J., Reizer A., Saier M.H. Jr. Microbiology 141:961-971(1995).

10   653. (Transpeptidase) Penicillin binding protein transpeptidase domain

The active site serine (residue 337 in Swiss;P14677) is conserved in all members of this family.

15   [1] Pares S, Mouz N, Petillot Y, Hakenbeck R, Dideberg O Nat Struct Biol 1996;3:284-289.

654. Trehalase signatures

Trehalase (EC 3.2.1.28) is the enzyme responsible for the degradation of the disaccharide
20   alpha, alpha-trehalose yielding two glucose subunits [1]. It is an enzyme found in a wide variety of organisms and whose sequence has been highly conserved throughout evolution. Two of the most highly conserved regions have been selected as signature patterns. The first pattern is located in the central section, the second one is in the C-terminal region.

Consensus pattern: P-G-G-R-F-x-E-x-Y-x-W-D-x-Y
25   Consensus pattern: Q-W-D-x-P-x-[GA]-W-[PAS]-P

[1] Kopp M., Mueller H., Holzer H. J. Biol. Chem. 268:4766-4774(1993).[2] Henrissat B.,
Bairoch A. Biochem. J. 293:781-788(1993).[E1]

30   655. Trehalose-6-phosphate synthase domain

OtsA (Trehalose-6-phosphate synthase) is homologous to regions
in the subunits of yeast trehalose-6-phosphate synthase/phosphate complex, [1].

[1] Kaasen I, McDougall J, Strom AR; Gene 1994;145:9-15.

656. Tropomyosins signature

Tropomyosins [1,2] are family of closely related proteins present in muscle and non-muscle cells. In striated muscle, tropomyosin mediate the interactions between the troponin complex
5   and actin so as to regulate muscle contraction. The role of tropomyosin in smooth muscle and non-muscle tissues is not clear. Tropomyosin is an alpha-helical protein that forms a coiled-coil dimer. Muscle isoforms of tropomyosin are characterized by having 284 amino acid residues and a highly conserved N-terminal region, whereas non-muscle forms are generally smaller and are heterogeneous in their N-terminal region. The signature pattern for
10   tropomyosins is based on a very conserved region in the C-terminal section of tropomyosins and which is present in both muscle and non-muscle forms.

Consensus pattern: L-K-E-A-E-x-R-A-E

[1] Smilie L.B. Trends Biochem. Sci. 4:151-155(1979).[2] McLeod A.R. BioEssays 6:208-212(1986).

15

657. Troponin

Troponin (Tn) contains three subunits, Ca2+ binding (TnC),

inhibitory (TnI), and tropomyosin binding (TnT). this Pfam contains
20   members of the TnT subunit.

Troponin is a complex of three proteins, Ca2+ binding (TnC),

inhibitory (TnI), and tropomyosin binding (TnT).

The troponin complex regulates Ca++ induced muscle contraction.

This family includes troponin T and troponin I. Troponin I
25   binds to actin and troponin T binds to tropomyosin.

Number of members: 81 [1]

Medline: 87144593

Structure of co-crystals of tropomyosin and troponin.

White SP, Cohen C, Phillips GN Jr;
30   Nature 1987;325:826-828.   [2]

Medline: 95155315

A direct regulatory role for troponin T and a dual role for

troponin C in the Ca2+ regulation of muscle contraction.

Potter JD, Sheng Z, Pan BS, Zhao J;

J Biol Chem 1995;270:2557-2562.

[3]Medline: 95324796

The troponin complex and regulation of muscle contraction.

Farah CS, Reinach FC;

5 FASEB J 1995;9:755-767.

658. (Tryp mucin) Mucin-like glycoprotein

10 This family of trypanosomal proteins resemble vertebrate mucins. The protein consists of

three regions. The N and C terminii are conserved between all members of the family,

whereas the central region is not well conserved and contains a large number of threonine

residues which can be glycosylated [1].

Indirect evidence suggested that these genes might encode the core protein of parasite

15 mucins, glycoproteins that were proposed to be involved in the interaction with, and invasion

of, mammalian host cells.

[1] Di Noia JM, Sanchez DO, Frasch AC; J Biol Chem 1995;270:24146-24149.

[2] Di Noia JM, D'Orso I, Aslund L, Sanchez DO, Frasch AC; J Biol Chem 1998;273:10843-

20 10850.

659. Aminoacyl-transfer RNA synthetases class-I signature (tRNA synt 1)

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino

25 acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In

prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA

synthetases, óne for each differentamino acid. In eukaryotes there are generally two

aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a

mitochondrial form. While all these enzymes have a common function, they are widely

30 diverse interms of subunit size and of quaternary structure. A few years ago it was found [2]

that several aminoacyl-tRNA synthetases share a region of similarity in their N-terminal

section, in particular the consensus tetrapeptide His-Ile-Gly-His ('HIGH') is very well

conserved. The 'HIGH' region has been shown [3] to be part of the adenylate binding site.

The 'HIGH' signature has been found in the aminoacyl-tRNA synthetases specific for

arginine, cysteine, glutamic acid, glutamine, isoleucine, leucine, methionine, tyrosine,
tryptophan, and valine. These aminoacyl-tRNA synthetases are referred to as class-I
synthetases [4,5,6] and seem to share the same tertiarystructure based on a Rossmann fold.
Consensus pattern: P-x(0,2)-[GSTAN]-[DENQGAPK]-x-[LIVMFP]-[HT]-[LIVMYAC]-G-

5   [HNTG]-[LIVMFYSTAGPC]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Webster T., Tsai H., Kula M.,
Mackie G.A., Schimmel P. Science 226:1315-1317(1984).[3] Brick P., Bhat T.N., Blow
D.M. J. Mol. Biol. 208:83-98(1988).[4] Delarue M., Moras D. BioEssays 15:675-
687(1993).[5] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[6] Nagel G.M., Doolittle

10   R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

660. Aminoacyl-transfer RNA synthetases class-I signature (tRNA synt 1b)

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino

15   acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In
prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA
synthetases, one for each different amino acid. In eukaryotes there are generally two
aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a
mitochondrial form. While all these enzymes have a common function, they are widely

20   diverse in terms of subunit size and of quaternary structure. A few years ago it was found [2]
that several aminoacyl-tRNA synthetases share a region of similarity in their N-terminal
section, in particular the consensus tetrapeptide His-Ile-Gly-His ('HIGH') is very well
conserved. The 'HIGH' region has been shown [3] to be part of the adenylate binding site.
The 'HIGH' signature has been found in the aminoacyl-tRNA synthetases specific

25   forarginine, cysteine, glutamic acid, glutamine, isoleucine, leucine, methionine, tyrosine,
tryptophan, and valine. These aminoacyl-tRNA synthetases are referred to as class-I
synthetases [4,5,6] and seem to share the same tertiary structure based on a Rossmann fold.
Consensus pattern: P-x(0,2)-[GSTAN]-[DENQGAPK]-x-[LIVMFP]-[HT]-[LIVMYAC]-G-
[HNTG]-[LIVMFYSTAGPC

30   [1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Webster T., Tsai H., Kula M.,
Mackie G.A., Schimmel P. Science 226:1315-1317(1984).[3] Brick P., Bhat T.N., Blow
D.M. J. Mol. Biol. 208:83-98(1988).[4] Delarue M., Moras D. BioEssays 15:675-
687(1993).[5] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[6] Nagel G.M., Doolittle
R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

661. (tRNA-synt 1C) tRNA synthetases class I (E and Q)

5   Other tRNA synthetase sub-families are too dissimilar to be included.

This family includes only glutamyl and glutaminyl tRNA synthetases.

In some organisms, a single glutamyl-tRNA synthetase aminoacylates both tRNA(Glu) and tRNA(Gln).

10 [1] Rath VL, Silvian LF, Beijer B, Sproat BS, Steitz TA; Structure 1998;6:439-449.

662. (tRNA-synt 1d) tRNA synthetases class I (R)

15 Other tRNA synthetase sub-families are too dissimilar to be included.

This family includes only arginyl tRNA synthetase.

663. Aminoacyl-transfer RNA synthetases class-II signatures (tRNA synt 2)

20 Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a

25 mitochondrial form. While all these enzymes have a common function, they are widely diverse interms of subunit size and of quaternary structure. The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is

30 different to the Rossmann fold observed for the class I synthetases [7].Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns have been derived from two of these regions.

Consensus pattern: [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE

Consensus pattern: [GSTALVF]-{DENQHRKP}-[GSTA]-[LIVMF]-[DE]-R-[LIVMF]-x-[LIVMSTAG]-[LIVMFY]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Delarue M., Moras D. BioEssays 15:675-687(1993).[3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[4] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991). [5] Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).[6] Cusack S. Biochimie 75:1077-1081(1993).[7] Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-255(1990).[8] Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).

664. Aminoacyl-transfer RNA synthetases class-I signature (tRNA synt 1e)
Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. A few years ago it was found [2] that several aminoacyl-tRNA synthetases share a region of similarity in their N-terminal section, in particular the consensus tetrapeptide His-Ile-Gly-His ('HIGH') is very well conserved. The 'HIGH' region has been shown [3] to be part of the adenylate binding site. The 'HIGH' signature has been found in the aminoacyl-tRNA synthetases specific forarginine, cysteine, glutamic acid, glutamine, isoleucine, leucine, methionine, tyrosine, tryptophan, and valine. These aminoacyl-tRNA synthetases are referred to as class-I synthetases [4,5,6] and seem to share the same tertiary structure based on a Rossmann fold.
Consensus pattern: P-x(0,2)-[GSTAN]-[DENQGAPK]-x-[LIVMFP]-[HT]-[LIVMYAC]-G-[HNTG]-[LIVMFYSTAGPC

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Webster T., Tsai H., Kula M., Mackie G.A., Schimmel P. Science 226:1315-1317(1984).[3] Brick P., Bhat T.N., Blow D.M. J. Mol. Biol. 208:83-98(1988).[4] Delarue M., Moras D. BioEssays 15:675-687(1993).[5] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[6] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

665. Aminoacyl-transfer RNA synthetases class-II signatures (tRNA synt 2b)

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse interms of subunit size and of quaternary structure. The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is different to the Rossmann fold observed for the class I synthetases [7].Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns have been derived from two of these regions.

Consensus pattern: [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE

Consensus pattern: [GSTALVF]-{DENQHRKP}-[GSTA]-[LIVMF]-[DE]-R-[LIVMF]-x-[LIVMSTAG]-[LIVMFY]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Delarue M., Moras D. BioEssays 15:675-687(1993).[3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[4] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991). [5] Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).[6] Cusack S. Biochimie 75:1077-1081(1993).[7] Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-255(1990).[8] Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).

666. Thaumatin family signature

Thaumatin [1] is an intensively sweet-tasting protein (100 000 times sweeter than sucrose on a molar basis) from Thaumatococcus daniellii, an African brush. The protein is made of about 200 residues and contains 8 disulfide bonds. A number of proteins have been found to be related to thaumatins. These protein are listed below (references are only provided for recently determined sequences). - A maize alpha-amylase/trypsin inhibitor. - Two tobacco pathogenesis-related proteins: PR-R major and minor forms, which are induced after

infection with viruses. - Salt-induced protein NP24 from tomato. - Osmotin, a salt-induced protein from tobacco. - Osmotin-like proteins OSML13, OSML15 and OSML81 from potato [2]. - P21, a leaf protein from soybean. - PWIR2, a leaf protein from wheat. - Zeamatin, a maize antifunal protein [3].The exact biological function of all these proteins is not yet

5   known. A conserved region that includes three cysteine residues known (in thaumatin) to be involved in disulfide bonds has been selected as a signature pattern.

```
       +--------------------------------------------------------------+ | +----------------+ | | ******* |
       | |
       xxCxxxxxxxxxxxxxxxxxCxxCxxCxCxxxxxxxxxxxxxxxCxxCxCxxxCxCxxCCxCxxxCxxxxxC
10     xxxCx | | | | | | | | || | | +---+ +-+ | +---+ +--++-+ | +------------------------+'C': conserved cysteine
       involved in a disulfide bond.'*': position of the pattern.
```

Consensus pattern: G-x-[GF]-x-C-x-T-[GA]-D-C-x(1,2)-G-x(2,3)-C

[1] Edens L., Heslinga L., Klok R., Ledeboer A.M., Maat J., Toonen M.Y., Visser C., Verrips C.T. Gene 18:1-12(1982).[2] Zhu B., Chen T.H.H., Li P.H. Plant Physiol. 108:929-

15   937(1995).[3] Malehorn D.E., Borgmeyer J.R., Smith C.E., Shah D.M.; Plant Physiol. 106:1471-1481(1994).
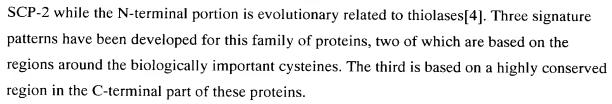
667. Thiolases signatures

20   Two different types of thiolase [1,2,3] are found both in eukaryotes and in prokaryotes: acetoacetyl-CoA thiolase (EC 2.3.1.9) and 3-ketoacyl-CoA thiolase(EC 2.3.1.16). 3-ketoacyl-CoA thiolase (also called thiolase I) has a broad chain-length specificity for its substrates and is involved in degradative pathways such as fatty acid beta-oxidation. Acetoacetyl-CoA thiolase (also called thiolase II) is specific for the thiolysis of acetoacetyl-CoA and involved

25   in biosynthetic pathways such as poly beta-hydroxybutyrate synthesisor steroid biogenesis. In eukaryotes, there are two forms of 3-ketoacyl-CoA thiolase: one located in the mitochondrion and the other in peroxisomes. There are two conserved cysteine residues important for thiolase activity. The first located in the N-terminal section of the enzymes is involved in the formation of an acyl-enzyme intermediate; the second located at the C-terminal extremity is

30   the active site base involved in deprotonation in the condensation reaction. Mammalian nonspecific lipid-transfer protein (nsL-TP) (also known as sterol carrier protein 2) is a protein which seems to exist in two different forms: a 14 Kd protein (SCP-2) and a larger 58 Kd protein (SCP-x). The former is found in the cytoplasm or the mitochondria and is involved in lipid transport; the latter is found in peroxisomes. The C-terminal part of SCP-x is identical to

SCP-2 while the N-terminal portion is evolutionary related to thiolases[4]. Three signature patterns have been developed for this family of proteins, two of which are based on the regions around the biologically important cysteines. The third is based on a highly conserved region in the C-terminal part of these proteins.

5   Consensus pattern: [LIVM]-[NST]-x(2)-C-[SAGLI]-[ST]-[SAG]-[LIVMFYNS]-x- [STAG]-[LIVM]-x(6)-[LIVM] [C is involved in formation of acyl-enzyme intermediate]
Consensus pattern: N-x(2)-G-G-x-[LIVM]-[SA]-x-G-H-P-x-[GA]-x-[ST]-G
Consensus pattern: [AG]-[LIVMA]-[STAGCLIVM]-[STAG]-[LIVMA]-C-x-[AG]-x-[AG]-x- [AG]-x-[SAG] [C is the active site residue]

10   [1] Peoples O.P., Sinskey A.J. J. Biol. Chem. 264:15293-15297(1989).[2] Yang S.-Y., Yang X.-Y.H., Healy-Louie G., Schulz H., Elzinga M. J. Biol. Chem. 265:10424-10429(1990).[3] Igual J.C., Gonzalez-Bosch C., Dopazo J., Perez-Ortin J.E. J. Mol. Evol. 35:147-155(1992).[4] Baker M.E., Billheimer J.T., Strauss J.F. III DNA Cell Biol. 10:695-698(1991).

15

668. Thioredoxin family active site

Thioredoxins [1 to 4] are small proteins of approximately one hundred amino-acid residues which participate in various redox reactions via the reversible oxidation of an active center disulfide bond. They exist in either a reduced form or an oxidized form where the two
20   cysteine residues are linked in an intramolecular disulfide bond. Thioredoxin is present in prokaryotes and eukaryotes and the sequence around the redox-active disulfide bond is wellconserved. Bacteriophage T4 also encodes for a thioredoxin but its primary structure is not homologous to bacterial, plant and vertebrate thioredoxins. A number of eukaryotic proteins contain domains evolutionary related tothioredoxin, all of them seem to be protein
25   disulphide isomerases (PDI). PDI(EC 5.3.4.1) [5,6,7] is an endoplasmic reticulum enzyme that catalyzes the rearrangement of disulfide bonds in various proteins. The various forms of PDI which are currently known are: - PDI major isozyme; a multifunctional protein that also function as the beta subunit of prolyl 4-hydroxylase (EC 1.14.11.2), as a component of oligosaccharyl transferase (EC 2.4.1.119), as thyroxine deiodinase (EC 3.8. 1.4), as
30   glutathione-insulin transhydrogenase (EC 1.8.4.2) and as a thyroid hormone-binding protein ! - ERp60 (ER-60; 58 Kd microsomal protein). ERp60 was originally thought to be a phosphoinositide-specific phospholipase C isozyme and later to be a protease. - ERp72. - P5.All PDI contains two or three (ERp72) copies of the thioredoxin domain. Bacterial proteins that act as thiol:disulfide interchange proteins thatallows disulfide bond formation in

some periplasmic proteins also contain a thioredoxin domain. These proteins are: - Escherichia coli dsbA (or prfA) and its orthologs in Vibrio cholerae (tcpG) and Haemophilus influenzae (por). - Escherichia coli dsbC (or xpRA) and its orthologs in Erwinia chrysanthemi and Haemophilus influenzae. - Escherichia coli dsbD (or dipZ) and its Haemophilus

5   influenzae ortholog. - Escherichia coli dsbE (or ccmG) and orthologs in Haemophilus influenzae, Rhodobacter capsulatus (helX), Rhiziobiacae (cycY and tlpA).

Consensus pattern: [LIVMF]-[LIVMSTA]-x-[LIVMFYC]-[FYWSTHE]-x(2)-[FYWGTN]-C- [GATPLVE]-[PHYWSTA]-C-x(6)-[LIVMFYWT] [The two C's form the redox-active bond]

10   [1] Holmgren A. Annu. Rev. Biochem. 54:237-271(1985).[2] Gleason F.K., Holmgren A. FEMS Microbiol. Rev. 54:271-297(1988).[3] Holmgren A. J. Biol. Chem. 264:13963-13966(1989).[4] Eklund H., Gleason F.K., Holmgren A. Proteins 11:13-28(1991).[5] Freedman R.B., Hawkins H.C., Murant S.J., Reid L. Biochem. Soc. Trans. 16:96-99(1988).[6] Kivirikko K.I., Myllyla R., Pihlajaniemi T. FASEB J. 3:1609-1617(1989).[7] Freedman

15   R.B., Hirst T.R., Tuite M.F. Trends Biochem. Sci. 19:331-336(1994).

669. (Transcript fac2) Transcription factor TFIIB repeat signature

In eukaryotes the initiation of transcription of protein encoding genes by polymerase II is

20   modulated by general and specific transcription factors. The general transcription factors operate through common promoters elements (such as the TATA box). At least seven different proteins associates to form the general transcription factors: TFIIA, -IIB, -IID, -IIE, -IIF, -IIG, and -IIH[1].Transcription factor IIB (TFIIB) plays a central role in the transcription of class II genes, it associates with a complex of TFIID-IIA bound to DNA (DA

25   complex) to form a ternary complex TFIID-IIA-IBB (DAB complex) which is then recognized by RNA polymerase II [2,3]. TFIIB is a protein of about 315 to 340amino acid residues which contains, in its C-terminal part an imperfect repeat of a domain of about 75 residues. This repeat could contribute an element of symmetry to the folded protein. The following proteins have been shown to be evolutionary related to TFIIB: - An archaebacterial

30   TFIIB homolog. In Pyrococcus woesei a previously undetected open reading frame has been shown [4] to be highly related to TFIIB. - Fungal transcription factor IIIB 70 Kd subunit (gene PCF4/TDS4/BRF1) [5]. This protein is a general activator of RNA polymerase III transcription and plays a role analogous to that of TFIIB in pol III transcription. The central

section of the repeated domain, which is the most conserved part of that domain has been selected as a signature pattern.

Consensus pattern: G-[KR]-x(3)-[STAGN]-x-[LIVMYA]-[GSTA](2)-[CSAV]-[LIVM]-[LIVMFY]-[LIVMA]-[GSA]-[STAC

5   [1] Weinmann R. Gene Expr. 2:81-91(1992).[2] Hawley D. Trends Biochem. Sci. 16:317-318(1991).[3] Ha I., Lane W.S., Reinberg D. Nature 352:689-695(1991).[4] Ouzounis C., Sander C. Cell 71:189-190(1992).[5] Khoo B., Brophy B., Jackson S.P. Genes Dev. 8:2879-2890(1994).

10

670. (transcritp fact) MADS-box domain signature and profile

A number of transcription factors contain a conserved domain of 56 amino-acid residues, sometimes known as the MADS-box domain [E1]. They are listed below: - Serum response factor (SRF) [1], a mammalian transcription factor that binds to the Serum Response Element

15   (SRE). This is a short sequence of dyad symmetry located 300 bp to the 5' end of the transcription initiation site of genes such as c-fos. - Mammalian myocyte-specific enhancer factors 2A to 2D (MEF2A to MEF2D). These proteins are transcription factor which binds specifically to the MEF2 element present in the regulatory regions of many muscle-specific genes. - Drosophila myocyte-specific enhancer factor 2 (MEF2). - Yeast GRM/PRTF protein

20   (gene MCM1) [2], a transcriptional regulator of mating-type-specific genes. - Yeast arginine metabolism regulation protein I (gene ARGR1 or ARG80). - Yeast transcription factor RLM1. - Yeast transcription factor SMP1. - Arabidopsis thaliana agamous protein (AG) [3], a probable transcription factor involved in regulating genes that determines stamen and carpel development in wild-type flowers. Mutations in the AG gene result in the replacement of the

25   stamens by petals and the carpels by a new flower. - Arabidopsis thaliana homeotic proteins Apetala1 (AP1), Apetala3 (AP3) and Pistillata (PI) which act locally to specify the identity of the floral meristem and to determine sepal and petal development [4]. - Antirrhinum majus and tobacco homeotic protein deficiens (DEFA) and globosa (GLO) [5]. Both proteins are transcription factors involved in the genetic control of flower development. Mutations in

30   DEFA or GLO cause the transformation of petals into sepals and of stamina into carpels. - Arabidopsis thaliana putative transcription factors AGL1 to AGL6 [6]. - Antirrhinum majus morphogenetic protein DEF H33 (squamosa).In SRF, the conserved domain has been shown [1] to be involved in DNA-binding and dimerization. A pattern that spans the complete length of the domain has been derived. The profile also spans the length of the MADS-box.

Consensus pattern: R-x-[RK]-x(5)-I-x-[DNGSK]-x(3)-[KR]-x(2)-T-[FY]-x-[RK](3)- x(2)-[LIVM]-x-K(2)-A-x-E-[LIVM]-[STA]-x-L-x(4)-[LIVM]-x- [LIVM](3)-x(6)-[LIVMF]-x(2)-[FY]

[1] Norman C., Runswick M., Pollock R., Treisman R. Cell 55:989-1003(1988).[2]

5   Passmore S., Maine G.T., Elble R., Christ C., Tye B.-K. J. Mol. Biol. 204:593-606(1988).[3]

Yanofsky M., Ma H., Bowman J., Drews G., Feldmann K.A., Meyerowitz E.M. Nature

346:35-39(1990).[4] Goto K., Meyerowitz E.M. Genes Dev. 8:1548-1560(1994).[5]

Troebner W., Ramirez L., Motte P., Hue I., Huijser P., Loennig W.-E., Saedler H., Sommer

H., Schwartz-Sommer Z. EMBO J. 11:4693-4704(1992).[6] Ma H., Yanofsky M.F.,

10   Meyerowitz E.M. Genes Dev. 5:484-495(1991).[E1]

671. Transketolase signatures

Transketolase (EC 2.2.1.1) (TK) catalyzes the reversible transfer of a two-carbon ketol unit

15   from xylulose 5-phosphate to an aldose receptor, such as ribose 5-phosphate, to form

sedoheptulose 7-phosphate and glyceraldehyde 3-phosphate. This enzyme, together with

transaldolase, provides a link between the glycolytic and pentose-phosphate pathways. TK

requires thiamin pyrophosphate as a cofactor. In most sources where TK has been purified, it

is a homodimer of approximately 70 Kd subunits. TK sequences from a variety of eukaryotic

20   and prokaryotic sources [1,2] show that the enzyme has been evolutionarily conserved. In the

peroxisomes of methylotrophic yeast Hansenula polymorpha, there is a highly related

enzyme, dihydroxy-acetone synthase (DHAS) (EC 2.2.1.3) (also known as formaldehyde

transketolase), which exhibits a very unusual specificity by including formaldehyde amongst

its substrates. 1-deoxyxylulose-5-phosphate synthase (DXP synthase) [3] is an enzyme so far

25   found in bacteria (gene dxs) and plants (gene CLA1) which catalyzes the thiamin

pyrophosphoate-dependent acyloin condensation reaction between carbon atoms 2 and 3 of

pyruvate and glyceraldehyde 3-phosphate to yield 1-deoxy-D- xylulose-5-phosphate (dxp), a

precursor in the biosynthetic pathway to isoprenoids, thiamin (vitamin B1), and pyridoxol

(vitamin B6). DXP synthase is evolutionary related to TK. Two regions of TK have been

30   selected as signature patterns. The first, located in the N-terminal section, contains a histidine

residue which appears to function inproton transfer during catalysis [4]. The second, located

in the central section, contains conserved acidic residues that are part of the active cleft and

may participate in substrate-binding [4].

527

Consensus pattern: R-x(3)-[LIVMTA]-[DENQSTHKF]-x(5,6)-[GSN]-G-H-[PLIVMF]-[GSTA]-x(2)-[LIMC]-[GS

Consensus pattern: G-[DEQGSA]-[DN]-G-[PAEQ]-[ST]-[HQ]-x-[PAGM]-[LIVMYAC]-[DEFYW]-x(2)-[STAP]-x(2)-[RGA]

5 [1] Abedinia M., Layfield R., Jones S.M., Nixon P.F., Mattick J.S. Biochem. Biophys. Res. Commun. 183:1159-1166(1992).[2] Fletcher T.S., Kwee I.L., Nakada T., Largman C., Martin B.M. Biochemistry 31:1892-1896(1992).[3] Sprenger G.A., Schorken U., Wiegert T., Grolle S., De Graaf A.A., Taylor S.V., Begley T.P., Bringer-Meyer S., Sahm H. Proc. Natl. Acad. Sci. U.S.A. 94:12857-12862(1997).[4] Lindqvist Y., Schneider G., Ermler U.,
10 Sundstroem M. EMBO J. 11:2373-2379(1992).

672. Transmembrane 4 family signature

Recently a number of eukaryotic cell surface antigens have been found to be evolutionary
15 related [1,2,3]. The proteins known to belong to this family are listed below: - Mammalian antigen CD9 (MIC3); A protein involved in platelet activation and aggregation. - Mammalian leukocyte antigen CD37, expressed on B lymphocytes. - Mammalian leukocyte antigen CD53 (OX-44), which may be involved in growth regulation in hematopoietic cells. - Mammalian lysosomal membrane protein CD63 (melanoma-associated antigen ME491; antigen AD1). -
20 Mammalian antigen CD81 (cell surface protein TAPA-1), which may play an important role in the regulation of lymphoma cell growth. - Mammalian antigen CD82 (protein R2; antigen C33; Kangai 1 (KAI1)), which associates with CD4 or CD8 and delivers costimulatory signals for the TCR/CD3 pathway. - Mammalian antigen CD151 (SFA-1; platelet-endothelial tetraspan antigen 3 (PETA-3)). - Mammalian cell surface glycoprotein A15 (TALLA-1;
25 MXS1). - Mammalian novel antigen 2 (NAG-2). - Human tumor-associated antigen CO-029. - Schistosoma mansoni and japonicum 23 Kd surface antigen (SM23 / SJ23).These proteins share the following characteristics: they all seem to be type III membrane proteins (type III proteins are integral membrane proteins that contain a N-terminal membrane-anchoring domain which is not cleaved during biosynthesis and which functions both as a translocation
30 signal and as a membrane anchor); they also contain three additional transmembrane regions, at least seven conserved cysteines residues, and are of approximately the same size (218 to 284 residues). These proteins are collectively know as the 'transmembrane 4 super family' (TM4) because they span the plasma membrane four times. A schematic diagram of the domain structure of these proteins isshown below. +-+-----+-------+----+-----+-----+-------------

----------------+-----+----+ | | TMa | Extra | TM2| Cyt | TM3 | Extracellular | TM4 | Cyt| +-+-----

+-------+----C-----C-----+--------CC---------C-----C---+-----C----+ *********Cyt : cytoplasmic

domain. TMa : transmembrane anchor.TM2 to TM4: transmembrane regions 2 to 4.'C' : conserved cysteine. '*' : position of the pattern.

5   A conserved region that includes two cysteines and seems to be located in a short cytoplasmic loop between two transmembrane domains has been selected as a signature for these proteins.

Consensus pattern: G-x(3)-[LIVMF]-x(2)-[GSA]-[LIVMF](2)-G-C-x-[GA]-[STA]- x(2)-[EG]-x(2)-[CWN]-[LIVM](2)

10   [1] Levy S., Nguyen V.Q., Andria M.L., Takahashi S. J. Biol. Chem. 266:14597-14602(1991).[2] Tomlinson M.G., Williams A.F., Wright M.D. Eur. J. Immunol. 23:136-40(1993).[3] Barclay A.N., Birkeland M.L., Brown M.H., Beyers A.D., Davis S.J., Somoza C., Williams A.F. The leucocyte antigen factbooks. Academic Press, London / San Diego, (1993).

15

673. Tryptophan synthase alpha chain signature

Tryptophan synthase catalyzes the last step in the biosynthesis of tryptophan: the conversion of indoleglycerol phosphate and serine, totryptophan and glyceraldehyde 3-phosphate [1,2]. It

20   has two functional domains: one for the aldol cleavage of indoleglycerol phosphate to indole andglyceraldehyde 3-phosphate and the other for the synthesis of tryptophan fromindole and serine. In bacteria and plants [3], each domain is found on a separate subunit (alpha and beta chains), while in fungi the two domains are fused together on a single multifunctional protein. A conserved region that contains three conserved acidic residues has been selected as a

25   signature pattern for the alpha chain. The first and the third acidic residues are believed to serve as proton donors/acceptors in the enzyme's catalytic mechanism.

Consensus pattern: [LIVM]-E-[LIVM]-G-x(2)-[FYC]-[ST]-[DE]-[PA]-[LIVMY]- [AGLI]-[DE]-G

[1] Crawford I.P. Annu. Rev. Microbiol. 43:567-600(1989).[2] Hyde C.C., Miles E.W.

30   Bio/Technology 8:27-32(1990).[3] Berlyn M.B., Last R.L., Fink G.R. Proc. Natl. Acad. Sci. U.S.A. 86:4604-4608(1989).

674. Tryptophan synthase beta chain pyridoxal-phosphate attachment site

Tryptophan synthase catalyzes the last step in the biosynthesis of tryptophan: the conversion of indoleglycerol phosphate and serine, totryptophan and glyceraldehyde 3-phosphate [1,2]. It has two functional domains: one for the aldol cleavage of indoleglycerol phosphate to indole andglyceraldehyde 3-phosphate and the other for the synthesis of tryptophan fromindole and

5   serine. In bacteria and plants [3], each domain is found on a separate subunit (alpha and beta chains), while in fungi the two domains arefused together on a single multifunctional protein. The beta chain of the enzyme requires pyridoxal-phosphate as a cofactor. The pyridoxal-phosphate group is attached to a lysine residue. The region around this lysine residue also contains two histidine residues which are part of the pyridoxal-phosphate binding site. The

10 signature pattern for the tryptophansynthase beta chain is derived from that conserved region.
    -Consensus pattern: [LIVM]-x-H-x-G-[STA]-H-K-x-N [K is the pyridoxal-P attachment site]
    [1] Crawford I.P. Annu. Rev. Microbiol. 43:567-600(1989).[2] Hyde C.C., Miles E.W. Bio/Technology 8:27-32(1990).[3] Berlyn M.B., Last R.L., Fink G.R. Proc. Natl. Acad. Sci. U.S.A. 86:4604-4608(1989).

15

    675. Serine proteases, trypsin family, active sites
    The catalytic activity of the serine proteases from the trypsin family is provided by a charge relay system involving an aspartic acid residue hydrogen-bonded to a histidine, which itself is

20 hydrogen-bonded to a serine. The sequences in the vicinity of the active site serine and histidine residues are well conserved in this family of proteases [1]. A partial list of proteases known to belong to the trypsin family is shown below. - Acrosin. - Blood coagulation factors VII, IX, X, XI and XII, thrombin, plasminogen, and protein C. - Cathepsin G. - Chymotrypsins. - Complement components C1r, C1s, C2, and complement factors B, D and

25 I. - Complement-activating component of RA-reactive factor. - Cytotoxic cell proteases (granzymes A to H). - Duodenase I. - Elastases 1, 2, 3A, 3B (protease E), leukocyte (medullasin). - Enterokinase (EC 3.4.21.9) (enteropeptidase). - Hepatocyte growth factor activator. - Hepsin. - Glandular (tissue) kallikreins (including EGF-binding protein types A, B, and C, NGF-gamma chain, gamma-renin, prostate specific antigen (PSA) and tonin). -

30 Plasma kallikrein. - Mast cell proteases (MCP) 1 (chymase) to 8. - Myeloblastin (proteinase 3) (Wegener's autoantigen). - Plasminogen activators (urokinase-type, and tissue-type). - Trypsins I, II, III, and IV. - Tryptases. - Snake venom proteases such as ancrod, batroxobin, cerastobin, flavoxobin, and protein C activator. - Collagenase from common cattle grub and collagenolytic protease from Atlantic sand fiddler crab. - Apolipoprotein(a). - Blood fluke

cercarial protease. - Drosophila trypsin like proteases: alpha, easter, snake-locus. - Drosophila protease stubble (gene sb). - Major mite fecal allergen Der p III. All the above proteins belong to family S1 in the classification of peptidases[2,E1] and originate from eukaryotic species. It should be noted thatbacterial proteases that belong to family S2A are similar

5 enough in the regions of the active site residues that they can be picked up by the same patterns. These proteases are listed below. - Achromobacter lyticus protease I. - Lysobacter alpha-lytic protease. - Streptogrisin A and B (Streptomyces proteases A and B). - Streptomyces griseus glutamyl endopeptidase II. - Streptomyces fradiae proteases 1 and 2. Consensus pattern: [LIVM]-[ST]-A-[STAG]-H-C [H is the active site residue]

10 Consensus pattern: [DNSTAGC]-[GSTAPIMVQH]-x(2)-G-[DE]-S-G-[GS]-[SAPHV]-[LIVMFYWH]-[LIVMFYSTANQH] [S is the active site residue]

[1] Brenner S. Nature 334:528-530(1988).[2] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).[E1]

15

676. (tsp) Thrombospondin type 1 domain

[1] Bork P; FEBS lett 1993;327:125-130.

20

677. Tubulin subunits alpha, beta, and gamma signature

Tubulins [1,2], the major constituent of microtubules are dimeric proteins which consist of two closely related subunits (alpha and beta). Tubulin binds two molecules of GTP at two different sites (N and E). At the E (Exchangeable) site, GTP is hydrolyzed during

25 incorporation into the microtubule. Near the E site is an invariant region rich in glycines which is found in both chains andwhich is now [3] said to control the access of the nucleotide to its binding site. A signature pattern was developed from this region. With the exception of the simple eukaryotes, most species express a variety of closely related alpha and beta isotypes. In most species there is a third member of the tubulin family: gamma tubulin.

30 Gamma tubulin is found at microtubule organizing centers (MTOC) such as the spindle poles or the centrosome, suggesting that it is involved in the minus-end nucleation of microtubule assembly [4].

Consensus pattern: [SAG]-G-G-T-G-[SA]-G

[1] Cleveland D.W., Sullivan K.F. Annu. Rev. Biochem. 54:331-365(1985).[2] Joshi H.C., Cleveland D.W. Cell Motil. Cytoskeleton 16:159-163(1990).[3] Hesse J., Thierauf M., Ponstingl H. J. Biol. Chem. 262:15472-15475(1987).[4] Joshi H.C. BioEssays 15:637-643(1993).

5

Tubulin-beta mRNA autoregulation signal

The stability of beta-tubulin mRNAs are autoregulated by their own translation product [1]. Unpolymerized tubulin subunits bind directly (or activate a factor(s) which binds co-translationally) to the nascent N-terminus of beta-tubulin. This binding is transduced through

10   the adjacent ribosomes to activatean RNAse that degrades the polysome-bound mRNA. The recognition element has been shown to be the first four amino acids of beta-tubulin: Met-Arg-Glu-Ile. Mutations to this sequence abolish the autoregulation effect (except for the replacement of Glu by Asp); transposition of this sequence to an internal region of a polypeptide also suppresses the autoregulatory effect.

15   Consensus pattern: <M-R-[DE]-[IL]

[1] Cleveland D.W. Trends Biochem. Sci. 13:339-343(1988).

678. (tRNA-synt 2c) Aminoacyl-transfer RNA synthetases class-II signatures. Aminoacyl-

20   tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a

25   mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is

30   different to the Rossmann fold observed for the class I synthetases [7].Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns have been derived from two of these regions.

Consensus pattern: [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]-

Consensus pattern: [GSTALVF]-{DENQHRKP}-[GSTA]-[LIVMF]-[DE]-R-[LIVMF]-x-[LIVMSTAG]-[LIVMFY]-

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Delarue M., Moras D. BioEssays 15:675-687(1993).[3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[4] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991). [5] Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).[6] Cusack S. Biochimie 75:1077-1081(1993).[7] Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-255(1990).[8] Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).

679. UBA-domain

The UBA-domain (ubiquitin associated domain) is a novel sequence motif found in several proteins having connections to ubiquitin and the ubiquitination pathway. The structure of the UBA domain consists of a compact three helix bundle [1]. Number of members: 84

[1] Structure of a human DNA repair protein UBA domain that interacts with HIV-1 Vpr. Dieckmann T, Withers-Ward ES, Jarosinski MA, Liu CF, Chen IS, Feigon J; Nat Struct Biol 1998;5:1042-1047.

680. UBX domain

Domain present in ubiquitin-regulatory proteins. Present in FAF1 and Shp1p.Number of members: 19

[1] The UBA domain: a sequence motif present in multiple enzyme classes of the ubiquitination pathway. Hofmann K, Bucher P; Trends Biochem Sci 1996;21:172-173.

681. (UCH) Ubiquitin carboxyl-terminal hydrolases family 1 cysteine active site

Ubiquitin carboxyl-terminal hydrolases (UCH) (deubiquitinating enzymes) [1,2] are thiol proteases that recognize and hydrolyze the peptide bond at the C-terminal glycine of ubiquitin. These enzymes are involved in the processing of poly-ubiquitin precursors as well as that of ubiquinated proteins. There are two distinct families of UCH. The first class consist

of enzymes ofabout 25 Kd and is currently represented by: - Mammalian isozymes L1 and L3. - Yeast YUH1. - Drosophila Uch.One of the active site residues of class-I UCH [3] is a cysteine. A signature pattern has been derived from the region around that residue.

Consensus pattern: Q-x(3)-N-[SA]-C-G-x(3)-[LIVM](2)-H-[SA]-[LIVM]-[SA] [C is the

5 active site residue

[1] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991).[2] D'andrea A., Pellman D. Crit. Rev. Biochem. Mol. Biol. 33:337-352(1998).[3] Johnston S.C., Larsen C.N., Cook W.J., Wilkinson K.D., Hill C.P. EMBO J. 16:3787-3796(1997).[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:461-486(1994).

10

682. Ubiquitin carboxyl-terminal hydrolases family 2 signatures (UCH-1)

Ubiquitin carboxyl-terminal hydrolases (UCH) (deubiquitinating enzymes) [1,2] are thiol proteases that recognize and hydrolyze the peptide bond at the C-terminal glycine of

15 ubiquitin. These enzymes are involved in the processing of poly-ubiquitin precursors as well as that of ubiquinated proteins. There are two distinct families of UCH. The second class consist of largeproteins (800 to 2000 residues) and is currently represented by: - Yeast UBP1, UBP2, UBP3, UBP4 (or DOA4/SSV7), UBP5, UBP7, UBP9, UBP10, UBP11, UBP12, UBP13, UBP14, UBP15 and UBP16. - Human tre-2. - Human isopeptidase T. - Human

20 isopeptidase T-3. - Mammalian Ode-1. - Mammalian Unp. - Mouse Dub-1. - Drosophila fat facets protein (gene faf). - Mammalian faf homolog. - Drosophila D-Ubp-64E. - Caenorhabditis elegans hypothetical protein R10E11.3. - Caenorhabditis elegans hypothetical protein K02C4.3.These proteins only share two regions of similarity. The first region containsa conserved cysteine which is probably implicated in the catalytic mechanism. The

25 second region contains two conserved histidines residues, one of which is also probably implicated in the catalytic mechanism. Signature patterns for both conserved regions have been developed.

Consensus pattern: G-[LIVMFY]-x(1,3)-[AGC]-[NASM]-x-C-[FYW]-[LIVMC]-[NST]-[SACV]-x-[LIVMS]-Q [C is the putative active site residue]

30 Consensus pattern: Y-x-L-x-[SAG]-[LIVMFT]-x(2)-H-x-G-x(4,5)-G-H-Y [The two H's are putative active site residues]

[1] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991).[2] D'andrea A., Pellman D. Crit. Rev. Biochem. Mol. Biol. 33:337-352(1998).[3] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:461-486(1994).

683. Ubiquitin carboxyl-terminal hydrolases family 2 signatures (UCH-2)

Ubiquitin carboxyl-terminal hydrolases (UCH) (deubiquitinating enzymes) [1,2] are thiol

5   proteases that recognize and hydrolyze the peptide bond at the C-terminal glycine of

ubiquitin. These enzymes are involved in the processing of poly-ubiquitin precursors as well

as that of ubiquinated proteins. There are two distinct families of UCH. The second class

consist of largeproteins (800 to 2000 residues) and is currently represented by: - Yeast UBP1,

UBP2, UBP3, UBP4 (or DOA4/SSV7), UBP5, UBP7, UBP9, UBP10, UBP11, UBP12,

10   UBP13, UBP14, UBP15 and UBP16. - Human tre-2. - Human isopeptidase T. - Human

isopeptidase T-3. - Mammalian Ode-1. - Mammalian Unp. - Mouse Dub-1. - Drosophila fat

facets protein (gene faf). - Mammalian faf homolog. - Drosophila D-Ubp-64E. -

Caenorhabditis elegans hypothetical protein R10E11.3. - Caenorhabditis elegans hypothetical

protein K02C4.3.These proteins only share two regions of similarity. The first region

15   containsa conserved cysteine which is probably implicated in the catalytic mechanism. The

second region contains two conserved histidines residues, one of which is also probably

implicated in the catalytic mechanism. Signature patterns for both conserved regions have

been developed.

Consensus pattern: G-[LIVMFY]-x(1,3)-[AGC]-[NASM]-x-C-[FYW]-[LIVMC]-[NST]-

20   [SACV]-x-[LIVMS]-Q [C is the putative active site residue]

Consensus pattern: Y-x-L-x-[SAG]-[LIVMFT]-x(2)-H-x-G-x(4,5)-G-H-Y [The two H's are

putative active site residues]

[1] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991).[2]

D'andrea A., Pellman D. Crit. Rev. Biochem. Mol. Biol. 33:337-352(1998).[3] Rawlings

25   N.D., Barrett A.J. Meth. Enzymol. 244:461-486(1994).

684. UDP-glycosyltransferases signature

UDP glycosyltransferases (UGT) are a superfamily of enzymes that catalyzes the addition of

30   the glycosyl group from a UTP-sugar to a small hydrophobic molecule. This family currently

consist of: - Mammalian UDP-glucoronosyl transferases (UDPGT) [1,2]. A large family of

membrane-bound microsomal enzymes which catalyze the transfer of glucuronic acid to a

wide variety of exogenous and endogenous lipophilic substrates. These enzymes are of major

importance in the detoxification and subsequent elimination of xenobiotics such as drugs and

carcinogens. - A large number of putative UDPGT from Caenorhabditis elegans. - Mammalian 2-hydroxyacylsphingosine 1-beta-galactosyltransferase [3] (also known as UDP-galactose-ceramide galactosyltransferase). This enzyme catalyzes the transfer of galactose to ceramide, a key enzymatic step in the biosynthesis of galactocerebrosides, which are

5   abundant sphingolipids of the myelin membrane of the central nervous system and peripheral nervous system. - Plants flavonol O(3)-glucosyltransferase. An enzyme [4] that catalyzes the transfer of glucose from UDP-glucose to a flavanol. This reaction is essential and one of the last steps in anthocyanin pigment biosynthesis. - Baculoviruses ecdysteroid UDP-glucosyltransferase (EC 2.4.1.-) [5] (egt). This enzyme catalyzes the transfer of glucose from

10   UDP-glucose to ectysteroids which are insect molting hormones. The expression of egt in the insect host interferes with the normal insect development by blocking the molting process. - Prokaryotic zeaxanthin glucosyl transferase (gene crtX), an enzyme involved in carotenoid biosynthesis and that catalyses the glycosylation reaction which converts zeaxanthin to zeaxanthin-beta- diglucoside. - Streptomyces macrolide glycosyltransferases [6]. These

15   enzymes specifically inactivates macrolide anitibiotics via 2'-O-glycosylation using UDP-glucose.These enzymes share a conserved domain of about 50 amino acid residues locatedin their C-terminal section and from which a pattern has been extracted todetect them. Consensus pattern: [FW]-x(2)-Q-x(2)-[LIVMYA]-[LIMV]-x(4,6)-[LVGAC]-[LVFYA]-[LIVMF]-[STAGCM]-[HNQ]-[STAGC]-G-x(2)-[STAG]-x(3)-[STAGL]- [LIVMFA]-x(4)-

20   [PQR]-[LIVMT]-x(3)-[PA]-x(3)-[DES]-[QEHN]

[1] Dutton G.J. (In) Glucoronidation of drugs and other compounds, Dutton G.J., Ed., pp 1-78, CRC Press, Boca Raton, (1980).[2] Burchell B., Nebert D.W., Nelson D.R., Bock K.W., Iyanagi T., Jansen P.L., Lancet D., Mulder G.J., Chowdhury J.R., Siest G., Tephly T.R., Mackenzie P.I. DNA Cell Biol. 10:487-494(1991).[3] Schulte S., Stoffel W. Proc. Natl.

25   Acad. Sci. U.S.A. 90:10265-10269(1993).[4] Furtek D., Schiefelbein J.W., Johnston F., Nelson O.E. Jr. Plant Mol. Biol. 11:473-481(1988).[5] O'Reilly D.R., Miller L.K. Science 245:1110-1112(1989).[6] Hernandez C., Olano C., Mendez C., Salas J.A. Gene 134:139-140(1993).

30

685. UDP-glucose/GDP-mannose dehydrogenase family

The UDP-glucose/GDP-mannose dehydrogenaseses are a small group of enzymes which possesses the ability to catlyze the NAD-dependent 2-fold oxidation of an alcholol to an acid without the release of an aldehyde intermediate [2]. Number of members: 55

[1] Purification and characterization of guanosine diphospho-D-mannose dehydrogenase. A key enzyme in the biosynthesis of alginate by Pseudomonas aeruginosa. Roychoudhury S, May TB, Gill JF, Singh SK, Feingold DS, Chakrabarty AM; J Biol Chem 1989;264:9380-9385. [2] Properties and kinetic analysis of UDP-glucose dehydrogenase

5   from group A streptococci. Irreversible inhibition by UDP-chloroacetol. Campbell RE, Sala RF, van de Rijn I, Tanner ME; J Biol Chem 1997;272:3416-3422.
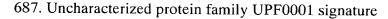

686. Uracil-DNA glycosylase signature

10   Uracil-DNA glycosylase (EC 3.2.2.-) (UNG) [1] is a DNA repair enzyme that excises uracil residues from DNA by cleaving the N-glycosylic bond. Uracil in DNA can arise as a result of misincorportation of dUMP residues by DNA polymerase or deamination of cytosine. The sequence of uracil-DNA glycosylase is extremely well conserved [2] in bacteria and eukaryotes as well as in herpes viruses. More distantly related uracil-DNA glycosylases are

15   also found in poxviruses [3].In eukaryotic cells, UNG activity is found in both the nucleus and the mitochondria. Human UNG1 protein is transported to both the mitochondria and the nucleus [4]. The N-terminal 77 amino acids of UNG1 seem to be required for mitochondrial localization [4], but the presence of a mitochondrial transitpeptide has not been directly demonstrated. As a signature for this type of enzyme, the most N-termina conserved region

20   has been selected. This region contains an aspartic acid residue which has been proposed, based on X-ray structures [5,6] to act as a general base in the catalytic mechanism.
Consensus pattern: [KR]-[LIV]-[LIVC]-[LIVM]-x-G-[QI]-D-P-Y [D is the active site residue]-

[1] Sancar A., Sancar G.B. Annu. Rev. Biochem. 57:29-67(1988).[2] Olsen L.C., Aasland
25   R., Wittwer C.U., Krokan H.E., Helland D.E. EMBO J. 8:3121-3125 (1989).[3] Upton C., Stuart D.T., McFadden G. Proc. Natl. Acad. Sci. U.S.A. 90:4518-4522(1993).[4] Slupphaug G., Markussen F.-H., Olsen L.C., Aasland R., Aarsaether N., Bakke O., Krokan H.E., Helland D.E. Nucleic Acids Res. 21:2579-2584(1993).[5] Savva R., McAuley-Hecht K., Brown T., Pearl L. Nature 373:487-493(1995).[6] Mol C.D., Arvai A.S., Slupphaug G., Kavli B.,
30   Alseth I., Krohan H.E., Tainer J.A. Cell 80:869-878(1995).[7] Muller S.J., Caradonna S. Biochim. Biophys. Acta 1088:197-207(1991).[8] Meyer-Siegler K., Mauro D.J., Seal G., Wurzer J., Deriel J.K., Sirover M.A. Proc. Natl. Acad. Sci. U.S.A. 88:8460-8464(1991).[9] Muller S.J., Caradonna S. J. Biol. Chem. 268:1310-1319(1993).[10] Barnes D.E., Lindahl T., Sedgwick B. Curr. Opin. Cell Biol. 5:424-433(1993).

687. Uncharacterized protein family UPF0001 signature

The following uncharacterized proteins have been shown [1] to share regions ofsimilarities: -

5   Yeast chromosome II hypothetical protein YBL036c. - Caenorhabditis elegans hypothetical

protein F09E5.8. - Bacillus subtilis hypothetical protein ylmE. - Escherichia coli hypothetical

protein yggS and HI0090, the corresponding Haemophilus influenzae protein. - Helicobacter

pylori hypothetical protein HP0395. - Mycobacterium tuberculosis hypothetical protein

MtCY270.20. - Synechocystis strain PCC 6803 hypothetical protein slr0556. - A

10   Pseudomonas aeruginosa hypothetical protein in pilT 5'region. - A Vibrio alginolyticus

hypothetical protein in pilT 5'region. These are proteins of from 25 to 30 Kd which contain a

number of conserved regions. The best conserved region which is located in the first third of

these proteins has been selected as a signature pattern.

Consensus pattern: [FW]-H-[FM]-[IV]-G-x-[LIV]-Q-x-[NKR]-K-x(3)-[LIV]

15   [1] Bairoch A., Rudd K.E. Unpublished observations (1996).

688. Uncharacterized protein family UPF0003 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -

20   Escherichia coli protein aefA. - Escherichia coli hypothetical protein yggB. - Escherichia coli

hypothetical protein yjeP and HI0195.1, the corresponding Haemophilus influenzae protein. -

Escherichia coli hypothetical protein ynaI. - Bacillus subtilis hypothetical protein yhdY. -

Helicobacter pylori hypothetical protein HP0415. - Synechocystis strain PCC 6803

hypothetical protein slr0639. - Archaeoglobus fulgidus hypothetical protein AF1546. -

25   Methanococcus jannaschii hypothetical protein MJ0170. - Methanococcus jannaschii

hypothetical protein MJ1143.The size of these proteins range from 30 to 120 Kd. They all

contain a number of transmembrane regions. The best conserved region which is located in

and just after the last potential transmembrane region has been selected as a signature

pattern,.

30   Consensus pattern: G-[STIF]-V-x(2)-[LIVM]-x(6)-[LIVMF]-x(3)-[DQ]-x(3)-[LIV]- x-[LIV]-

P-N-x(2)-[LIVMF]-[LIVFSTA]-x(5)-N

[1] Bairoch A. Unpublished observations (1997).

689. Uncharacterized protein family UPF0004 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Escherichia coli hypothetical protein yliG. - Escherichia coli hypothetical protein yleA and HI0019, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical

5   protein yqeV. - Helicobacter pylori hypothetical protein HP0269. - Helicobacter pylori hypothetical protein HP0285. - Mycoplasma iowae hypothetical protein in 16S RNA 5'region. - Mycobacterium leprae hypothetical protein B2235_C2_195. - Pseudomonas aeruginosa hypothetical protein in hemL 3'region. - Synechocystis strain PCC 6803 hypothetical protein slr0082. - Synechocystis strain PCC 6803 hypothetical protein sll0996. -

10   Methanococcus jannaschii hypothetical protein MJ0865. - Methanococcus jannaschii hypothetical protein MJ0867. - Caenorhabditis elegans hypothetical protein F25B5.5.The size of these proteins range from 47 to 61 Kd. They contain six conserved cysteines, three of which are clustered in a region that can be used as asignature pattern.

Consensus pattern: [LIVM]-x-[LIVMT]-x(2)-G-C-x(3)-C-[STAN]-[FY]-C-x-[LIVM]- x(4)-

15   G

[1] Bairoch A. Unpublished observations (1997).

690. Uncharacterized protein family UPF0005 signature

20   The following proteins seems to be evolutionary related [1]: - Mammalian protein TEGT (Testis Enhanced Gene Transcript). - Escherichia coli hypothetical protein yccA and HI0044, the corresponding Haemophilus influenzae protein. - A probable Pseudomonas aeruginosa ortholog of yccA. These are proteins of about 25 Kd which seem to contain seven transmembranedomains. A signature pattern that corresponds to a region that starts with the

25   beginning of the third transmembrane domain and ends in the middle of the fourth one has been developed.

Consensus pattern: G-[LIVM](2)-[SA]-x(5,8)-G-x(2)-[LIVM]-G-P-x-L-x(4)-[SAG]- x(4,6)-[LIVM](2)-x(2)-A-x(3)-T-A-[LIVM](2)-F

[1] Walter L., Marynen P., Szpirer J., Levan G., Guenther E. Genomics 28:301-304(1995).

30

691. Uncharacterized protein family UPF0006 signatures

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Yeast chromosome II hypothetical protein YBL055c. - Escherichia coli hypothetical protein

ycfH and HI0454, the corresponding Haemophilus influenzae protein. - Escherichia coli
hypothetical protein yigW. - Escherichia coli hypothetical protein yjjV and HI0081, the
corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein yabD.
- Haemophilus influenzae hypothetical protein HI1664. - Mycoplasma genitalium

5   hypothetical protein MG009. These are proteins of from 24 to 47 Kd which contain a number
of conserved regions. They can be picked up in the database by the following patterns.
Consensus pattern: [LIVMFY](2)-D-[STA]-H-x-H-[LIVMF]-[DN
Consensus pattern: P-[LIVM]-x-[LIVM]-H-x-R-x-[TA]-x-[DE
Consensus pattern: [LVSA]-[LIVA]-x(2)-[LIVM]-[PS]-x(3)-L-[LIVM]-[LIVMS]-E-T- D-x-

10   P

[1] Bairoch A., Rudd K.E. Unpublished observations (1995).

692. Uncharacterized protein family UPF0007 signature

15   The following proteins seems to be evolutionary related [1]: - Escherichia coli hypothetical
protein ygbP and HI0672, the corresponding Haemophilus influenzae protein. - Bacillus
subtilis hypothetical protein yacM. - Mycobacterium tuberculosis hypothetical protein
MtCY06G11.29c. - Synechocystis strain PCC 6803 hypothetical protein slr0951. - A
Rhodobacter capsulatus hypothetical protein in nifR3 5'region. Except for the Rhodobacter

20   protein which contains a C-terminal extension, all these proteins have from 225 to 236 amino
acids. They are hydrophilic proteins that can be picked up in the database by the following
pattern.
Consensus pattern: V-L-[IV]-H-D-[GA]-A-R

[1] Bairoch A. Unpublished observations (1997).

25

693. Uncharacterized protein family UPF0015 signature
The following uncharacterized proteins have been shown [1] to share regions of similarities: -
Yeast chromosome II hypothetical protein YBR002c. - Yeast chromosome XIII hypothetical

30   protein YMR101c. - Escherichia coli hypothetical protein yaeU and HI0920, the
corresponding Haemophilus influenzae protein. - Helicobacter pylori hypothetical protein
HP1221. - Mycobacterium leprae hypothetical protein B1937_F2_65. - A Corynebacterium
glutamicum hypothetical protein in aroF 3'region. - A Streptomyces fradiae hypothetical
protein in transposon Tn4556. - Synechocystis strain PCC 6803 hypothetical protein sll0505.

- Methanococcus jannaschii hypothetical protein MJ1372.These are proteins of about 26 to 40 Kd whose central region is well conserved. They can be picked up in the database by the following pattern.

Consensus pattern: [DE]-[LIVMF](3)-R-T-[SG]-G-x(2)-R-x-S-x-[FY]-[LIVM](2)-W-Q-

5   [1] Wolfe K.H., Lohan A.J.E. Yeast 10:S41-S46(1994).

694. Uncharacterized protein family UPF0016 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -

10   Yeast hypothetical protein YBR187w. - Fission yeast hypothetical protein SpAC17G8.08c. - Mouse protein pFT27. - Synechocystis strain PCC 6803 hypothetical protein sll0615. These are hydrophobic proteins of 200 to 320 amino acids that seem to contain six or seven transmembrane domains. A conserved region which seems, in the eukaryotic proteins of this family, to directly follow the second transmembrane domain has been selected as a signature

15   pattern.

Consensus pattern: E-[LIVM]-G-D-K-T-F-[LIVMF](2)-A-

[1] Bairoch A. Unpublished observations (1996).

20   695. Uncharacterized protein family UPF0021 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Yeast chromosome VII hypothetical protein YGL211w. - Dictyostelium discoideum protein veg136. - Methanococcus jannaschii hypothetical proteins MJ1157 and MJ1478.These are proteins of from 300 to 36o residues. They can be picked up in thedatabase by the following

25   pattern which is located in their N-terminalsection.

Consensus pattern: C-K-x(2)-F-x(4)-E-x(22,23)-S-G-G-K-D

[1] Bairoch A. Unpublished observations (1997).

30   696. Uncharacterized protein family UPF0023 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Mouse protein 22A3. - Yeast chromosome XII hypothetical protein YLR022c. - Caenorhabditis elegans hypothetical protein W06E11.4. - Methanococcus jannaschii

hypothetical protein MJ0592.These are hydrophilic proteins of about 30 Kd. They can be picked up in the database by the following pattern.

Consensus pattern: D-x-D-E-[LIV]-L-x(4)-V-F-x(3)-S-K-G-

[1] Bairoch A. Unpublished observations (1997).

5

697. Uncharacterized protein family UPF0024 signature. The following uncharacterized proteins have been shown [1] to share regions of similarities: - Escherichia coli hypothetical protein ygbO and HI0701, the corresponding Haemophilus influenzae protein. - Helicobacter pylori hypothetical protein HP0926. - Yeast chromosome XV hypothetical protein YOR243c. - Caenorhabditis elegans hypothetical protein B0024.11. - Methanococcus jannaschii hypothetical proteins MJ0588 and MJ1364.These are hydrophilic proteins of from 39 to 77 Kd. They can be picked up in the database by the following pattern.

15   Consensus pattern: G-x-K-D-[KR]-x-A-[LV]-T-x-Q-x-[LIVF]-[SGC]-

[1] Bairoch A. Unpublished observations (1997).

20   698. Uncharacterized protein family UPF0025 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Escherichia coli hypothetical protein yfcE. - Bacillus subtilis hypothetical protein ysnB. - Mycoplasma genitalium and pneumoniae hypothetical protein MG207. - Methanococcus jannaschii hypothetical proteins MJ0623 and MJ0936. These are hydrophilic proteins of

25   about 20 Kd. They can be picked up in thedatabase by the following pattern.

Consensus pattern: D-V-[LIV]-x(2)-G-H-[ST]-H-x(12)-[LIVMF]-N-P-G

[1] Bairoch A. Unpublished observations (1997).

30   699. Uncharacterized protein family UPF0029 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Yeast chromosome III hypothetical protein YCR59c. - Yeast chromosome IV hypothetical protein YDL177C. - Escherichia coli hypothetical protein yigZ and HI0722, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein yvyE.

- A Thermus aquaticus hypothetical protein in pol 5'region. These proteins can be picked up in the database by the following pattern.

Consensus pattern: G-x(2)-[LIVM](2)-x(2)-[LIVM]-x(4)-[LIVM]-x(5)-[LIVM](2)-x- R-[FYW](2)-G-G-x(2)-[LIVM]-G

5 [1] Koonin E.V., Bork P., Sander C. EMBO J. 13:493-503(1994).

700. Uncharacterized protein family UPF0030 signature

The following uncharacterized proteins have been shown [1] to be highly similar: - Yeast

10 chromosome VI hypothetical protein YFL060c. - Yeast chromosome XIII hypothetical protein YMR095c. - Yeast chromosome XIV hypothetical protein YNL334c. - Bacillus subtilis hypothetical protein yaaE. - Haemophilus influenzae hypothetical protein HI1648. - Methanococcus jannaschii hypothetical protein MJ1661.These are hydrophilic proteins of about 19 to 25 Kd. They can be picked up inthe database by the following pattern.

15 Consensus pattern: [GA]-L-I-[LIV]-P-G-G-E-S-T-[STA]

[1] Bairoch A. Unpublished observations (1997).

701. Uncharacterized protein family UPF0032 signature

20 The following uncharacterized proteins have been shown [1] to share regions of similarities: - Escherichia coli hypothetical protein yigU and HI0188, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein ycbT. - Mycobacterium tuberculosis hypothetical protein MtCY49.33c and U2126A, the corresponding Mycobacterium leprae protein. - Synechocystis strain PCC 6803 hypothetical protein sll0194.

25 - Odontella sinensis and Porphyra purpurea chlroplast hypothetical protein ycf43.These proteins have from 245 to 317 amino acids and seem to contain at least six or seven transmembrane regions. A conserved region located in the central section of these proteins has been developed as a signature pattern,.

Consensus pattern: Y-x(2)-F-[LIVMA](2)-x-L-x(4)-G-x(2)-F-[EQ]-[LIVMF]-P- [LIVM] –

30 [1] Bairoch A., Rudd K.E. Unpublished observations (1996).

702. Uncharacterized protein family UPF0034 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Escherichia coli hypothetical protein yhdG and HI0979, the corresponding Haemophilus influenzae protein. - Escherichia coli hypothetical protein yjbN and HI0634, the corresponding Haemophilus influenzae protein. - Escherichia coli hypothetical protein yohI

5 and HI0270, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein yacF. - Rhodobacter capsulatus protein nifR3 and related proteins in Azospirillum brasilense and Rhizobium leguminosarum. - Synechocystis strain PCC 6803 hypothetical protein slr0644. - Synechocystis strain PCC 6803 hypothetical protein sll0926. - Caenorhabditis elegans hypothetical protein C45G9.2. - Yeast protein SMM1. - Yeast

10 hypothetical protein YLR401c. - Yeast hypothetical protein YLR405w. - Yeast hypothetical protein YML080w. Although it has been proposed [2] that Rhodobacter capsulatus nifR3 is a transcriptional regulatory protein, it is believed that these proteins constitute a family of enzymes whose active site could include a conserved cysteine which has been used as the central part of a signature pattern.

15 Consensus pattern: [LIVM]-[DNG]-[LIVM]-N-x-G-C-P-x(3)-[LIVMASQ]-x(5)-G-[SAC]
[1] Bairoch A., Rudd K.E. Unpublished observations (1995).[2] Foster-Hartnett D., Cullen P.J., Gabbert K.K., Kranz R.G. Mol. Microbiol. 8:903-914(1993).

20 703. Uncharacterized protein family UPF0038 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Escherichia coli hypothetical protein yacE and HI0890, the corresponding Haemophilus influenzae protein. - Mycobacterium tuberculosis hypothetical protein MtCY01B2.23 and O410, the corresponding Mycobacterium leprae protein. - Synechocystis strain PCC 6803

25 hypothetical protein slr0553. - Other hypothetical proteins from Aeromonas hydrophila, Bacteroides nodosus, Neisseria gonorrhoeae, Pseudomonas putida, Thermus thermophilus and Xanthomonas campestris. - Human hypothetical protein pOV-2. - Yeast hypothetical protein YDR196C. - Caenorhabditis elegans hypothetical protein T05G5.5.These proteins all contain, in their N-terminal extremity, an ATP/GTP-binding motif 'A' (P-loop) (see

30 <PDOC00017>). The size of these proteins range from 200 to 290 residues (with the exception of the Mycobacterial sequences which are are 410 residues long). A conseved region some 50 residues away from the ATP-binding P-loop has been developed as a signature pattern.

Consensus pattern: G-x-[LI]-x-R-x(2)-L-x(4)-F-x(8)-[LIV]-x(5)-P-x-[LIV] –

[1] Rudd K.E., Bairoch A. Unpublished observations (1997).

704. Ubiquitin-conjugating enzymes active site

Ubiquitin-conjugating enzymes (UBC or E2 enzymes) [1,2,3] catalyze the covalent attachment of ubiquitin to target proteins. An activatedubiquitin moiety is transferred from an ubiquitin-activating enzyme (E1) to E2which later ligates ubiquitin directly to substrate proteins with or without the assistance of 'N-end' recognizing proteins (E3). In most species there are many forms of UBC (at least 9 in yeast) which are implicated in diverse cellular functions. A cysteine residue is required for ubiquitin-thiolester formation. There is a single conserved cysteine in UBC's and the region around that residue isconserved in the sequence of known UBC isozymes. That region has been used as a signature pattern.

Consensus pattern: [FYWLSP]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x-[LIV]-C-[LIV]-x- [LIV] [C is the active site residue]

[1] Jentsch S., Seufert W., Sommer T., Reins H.-A. Trends Biochem. Sci. 15:195-198(1990).[2] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991).[3] Hershko A. Trends Biochem. Sci. 16:265-268(1991).

705. Uroporphyrinogen decarboxylase signatures

Uroporphyrinogen decarboxylase (URO-D), the fifth enzyme of the heme biosynthetic pathway, catalyzes the sequential decarboxylation of the four acetyl side chains of uroporphyrinogen to yield coproporphyrinogen [1].URO-D deficiency is responsible for the Human genetic diseases familialporphyria cutanea tarda (fPCT) and hepatoerythropoietic porphyria (HEP).The sequence of URO-D has been well conserved throughout evolution. The best conserved region is located in the N-terminal section; it contains a perfectlyconserved hexapeptide. There are two arginine residues in this hexapeptide which could be involved in the binding, via salt bridges, to the carboxylgroups of the propionate side chains of the substrate. This region has been used as a signature pattern. A second signature pattern is based on a another well conserved region which is located in the central section of the protein.

Consensus pattern: P-x-W-x-M-R-Q-A-G-R

Consensus pattern: G-F-[STAGCV]-[STAGC]-x-P-[FYW]-T-[LV]-x(2)-Y-x(2)-[AE]- [GK]

[1] Garey J.R., Labbe-Bois R., Chelstowska A., Rytka J., Harrison L., Kushner J., Labbe P. Eur. J. Biochem. 205:1011-1016(1992).

5   706. ubiE/COQ5 methyltransferase family signatures

The following methyltransferases have been shown [1] to share regions of similarities: - Escherichia coli ubiE, which is involved in both ubiquinone and menaquinone biosynthesis and which catalyzes the S-adenosylmethionine dependent methylation of 2-polyprenyl-6-methoxy-1,4-benzoquinol into 2-polyprenyl-3- methyl-6-methoxy-1,4-benzoquinol and of

10   demethylmenaquinol into menaquinol. - Yeast COQ5, a ubiquinone biosynthesis methlytransferase. - Bacillus subtilis spore germination protein C2 (gene: gercB or gerC2), a probable menaquinone biosynthesis methlytransferase. - Lactococcus lactis gerC2 homolog. - Caenorhabditis elegans hypothetical protein ZK652.9. - Leishmania donovani amastigote-specific protein A41.These are hydrophilic proteins of about 30 Kd (except for ZK652.9

15   which is 65Kd). They can be picked up in the database by the following patterns.

Consensus pattern: Y-D-x-M-N-x(2)-[LIVM]-S-x(3)-H-x(2)-W

Consensus pattern: R-V-[LIVM]-K-[PV]-G-G-x-[LIVMF]-x(2)-[LIVM]-E-x-S

[1] Lee P.T., Hsu A.Y., Ha H.T., Clarke C.F. J. Bacteriol. 179:1748-1754(1997).

20

707. Uricase signature

Uricase (urate oxidase) [1] is the peroxisomal enzyme responsible for the degradation of urate into allantoin. Some species, like primates and birds, have lost the gene for uricase and are therefore unable to degradeurate. Uricase is a protein of 300 to 400 amino acids. A highly

25   conserved region located in the central part of the sequence has been used as a signature pattern.

Consensus pattern: [LV]-x-[LV]-[LIV]-K-[STV]-[ST]-x-[SN]-x-F-x(2)-[FY]-x(4)- [FY]-x(2)-L-x(5)-R

[1] Motojima K., Kanaya S., Goto S. J. Biol. Chem. 263:16677-16681(1988).

30

708. Universal stress protein family (Usp)

By a wide range of stress conditions members of the Usp family are predicted to be related to the MADS-box proteins transcript_fact and bind to DNA [2]. Number of members: 39

5   [1] Expression and role of the universal stress protein, UspA, of Escherichia coli during growth arrest. Nystrom T, Neidhardt FC; Mol Microbiol 1994; 11:537-544.

[2] Sequence analysis of eukaryotic developmental proteins: ancient and novel domains. Mushegian AR, Koonin EV; Genetics 1996; 144:817-828.

10

709. Ubiquitin domain signature and profile

Ubiquitin [1,2,3] is a protein of seventy six amino acid residues, found in all eukaryotic cells and whose sequence is extremely well conserved from protozoan to vertebrates. It plays a key role in a variety of cellular processes, such as ATP-dependent selective degradation of

15   cellular proteins,maintenance of chromatin structure, regulation of gene expression, stress response and ribosome biogenesis. In most species, there are many genes coding for ubiquitin. However they can be classified into two classes. The first class produces polyubiquitin molecules consisting of exact head to tail repeats of ubiquitin. The number of repeats is variable (up to twelve in a Xenopus gene). In the majority of polyubiquitin

20   precursors, there is a final amino-acid after the last repeat. The second class of genes produces precursor proteins consisting of a single copy of ubiquitin fused to a C-terminal extension protein (CEP). There are two types of CEP proteins and both seem to be ribosomal proteins. Ubiquitin is a globular protein, the last four C-terminal residues (Leu-Arg- Gly-Gly) extending from the compact structure to form a 'tail', important for its function. The latter is

25   mediated by the covalent conjugation of ubiquitin to target proteins, by an isopeptide linkage between the C-terminal glycine and the epsilon amino group of lysine residues in the target proteins. There are a number of proteins which are evolutionary related to ubiquitin: - Ubiquitin-like proteins from baculoviruses as well as in some strains of bovine viral diarrhea viruses (BVDV). These proteins are highly similar to their eukaryotic counterparts. -

30   Mammalian protein GDX [4]. GDX is composed of two domains, a N-terminal ubiquitin-like domain of 74 residues and a C-terminal domain of 83 residues with some similarity with the thyroglobulin hormonogenic site. - Mammalian protein FAU [5]. FAU is a fusion protein which consist of a N-terminal ubiquitin-like protein of 74 residues fused to ribosomal protein S30. - Mouse protein NEDD-8 [6], a ubiquitin-like protein of 81 residues. - Human protein

BAT3, a large fusion protein of 1132 residues that contains a N-terminal ubiquitin-like domain. - Caenorhabditis elegans protein ubl-1 [7]. Ubl-1 is a fusion protein which consist of a N-terminal ubiquitin-like protein of 70 residues fused to ribosomal protein S27A. - Yeast DNA repair protein RAD23 [8]. RAD23 contains a N-terminal domain that seems to be

5   distantly, yet significantly, related to ubiquitin. - Mammalian RAD23-related proteins RAD23A and RAD23B. - Mammalian BCL-2 binding athanogene-1 (BAG-1). BAG-1 is a protein of 274 residues that contains a central ubiquitin-like domain. - Human spliceosome associated protein 114 (SAP 114 or SF3A120). - Yeast protein DSK2, a protein involved in spindle pole body duplication and which contains a N-terminal ubiquitin-like domain. -

10   Human protein CKAP1/TFCB, Schizosaccharomyces pombe protein alp11 and Caenorhabditis elegans hypothetical protein F53F4.3. These proteins contain a N-terminal ubiquitin domain and a C-terminal CAP-Gly domain. - Schizosaccharomyces pombe hypothetical protein SpAC26A3.16. This protein contains a N-terminal ubiquitin domain. - Yeast protein SMT3. - Human ubiquitin-like proteins SMT3A and SMT3B. - Human

15   ubiquitin-like protein SMT3C (also known as PIC1; Ubl1, Sumo-1; Gmp-1 or Sentrin). This protein is involved in targeting ranGAP1 to the nuclear pore complex protein ranBP2. - SMT3-like proteins in plants and Caenorhabditis elegans. To identify ubiquitin and related proteins, a pattern has been developed based on conserved positions in the central section of the sequence. A profile was also developed that spans the complete length of the ubiquitin

20   domain.

Consensus pattern: K-x(2)-[LIVM]-x-[DESAK]-x(3)-[LIVM]-[PA]-x(3)-Q-x-[LIVM]-[LIVMC]-[LIVMFY]-x-G-x(4)-[DE]

[1] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991).[2] Monia B.P., Ecker D.J., Croke S.T. Bio/Technology 8:209-215(1990).[3] Finley D.,

25   Varshavsky A. Trends Biochem. Sci. 10:343-347(1985).[4] Filippi M., Tribioli C., Toniolo D. Genomics 7:453-457(1990).[5] Olvera J., Wool I.G. J. Biol. Chem. 268:17967-17974(1993).[6] Kumar S., Yoshida Y., Noda M. Biochem. Biophys. Res. Commun. 195:393-399(1993).[7] Jones D., Candido E.P. J. Biol. Chem. 268:19545-19551(1993).[8] Melnick L., Sherman F. J. Mol. Biol. 233:372-388(1993).

30

710. VHS domain

    Domain present in VPS-27, Hrs and STAM. Number of members: 27

711. Vinculin family signatures

Vinculin [1] is a eukaryotic protein that seems to be involved in the attachment of the actin-based microfilaments to the plasma membrane. Vinculinis located at the cytoplasmic side of focal contacts or adhesion plaques. In addition to actin, vinculin interacts with other structural proteins such as talin and alpha-actinins. Vinculin is a large protein of 116 Kd (about a 1000 residues). Structurally the protein consists of an acidic N-terminal domain of about 90 Kd separated from a basic C-terminal domain of about 25 Kd by a proline-rich region of about 50 residues. The central part of the N-terminal domain consists of avariable number (3 in vertebrates, 2 in Caenorhabditis elegans) of repeats of a 110 amino acids domain. Catenins [2] are proteins that associate with the cytoplasmic domain of avariety of cadherins. The association of catenins to cadherins produces a complex which is linked to the actin filament network, and which seems to be of primary importance for cadherins cell-adhesion properties. Three different types of catenins seem to exist: alpha, beta, and gamma. Alpha-catenins are proteins of about 100 Kd which are evolutionary related to vinculin. Interm of their structure the most significant differences are the absence, inalpha-catenin, of the repeated domain and of the proline-rich segment. Two signature patterns for this family of proteins have been devolped. The first pattern is located in the N-terminal section of both vinculin and alpha-catenins and is part, in vinculin, of a domain that seems to be involved with the interaction with talin. The second pattern is based on a conserved regionin the N-terminal part of the repeated domain of vinculin.

Consensus pattern: [KR]-x-[LIVMF]-x(3)-[LIVMA]-x(2)-[LIVM]-x(6)-R-Q-Q-E-L

Consensus pattern: [LIVM]-x-[QA]-A-x(2)-W-[IL]-x-[DN]-P

[1] Otto J.J. Cell Motil. Cytoskeleton 16:1-6(1990).[2] Herrenknecht K., Ozawa M., Eckerskorn C., Lottspeich F., Lenter M., Kemler R. Proc. Natl. Acad. Sci. U.S.A. 88:9156-9160(1991).

712. (Vitellogenin N) Lipoprotein amino terminal region

This family contains regions from: Vitellogenin, Microsomal triglyceride transfer protein and apolipoprotein B-100. These proteins are all involved in lipid transport [1]. This family contains the LV1n chain from lipovitellin, that contains two structural domains. Number of members: 33

[1] The structural basis of lipid interactions in lipovitellin, a soluble lipoprotein.

Anderson TA, Levitt DG, Banaszak LJ Structure 1998;6:895-909.

5   713. (VMSA) Major surface antigen from hepadnavirus

714. ssDNA binding protein (Viral DNA bp)

This protein is found in herpesviruses and is needed for

10   replication.

715. (Votage CLC) Voltage gated chloride channels

15   This family of ion channels contains 10 or 12 transmembrane helices. Each protein forms a

single pore. It has been shown that some members of this family form homodimers. These

proteins contain two CBS domains.

[1] Schmidt-Rose T, Jentsch TJ; J Biol Chem 1997;272:20515-20521.

20   [2] Zhang J, George AL Jr, Griggs RC, Fouad GT, Roberts J, Kwiecinski H, Connolly AM,

Ptacek LJ; Neurology 1996;47:993-998.

716. von Willebrand factor type A domain (vwa)

25   More von Willebrand factor type A domains? Sequence

similarities with malaria thrombospondin-related

anonymous protein, dihydropyridine-sensitive calcium

channel and inter-alpha-trypsin inhibitor.

Bork P, Rohde K;

30   Biochem J 1991;279:908-911.

1. RUGGERI, Z.M. and WARE, J.

von Willebrand factor.

FASEB J. 7 308-316 (1993).

2. COLOMBATTI, A., BONALDO, P. and DOLIANA, R.

Type A modules: interacting domains found in several non-fibrillar

collagens and in other extracellular matrix proteins.

5   MATRIX 13 297-306 (1993).

3. PERKINS, S.J., SMITH, K.F., WILLIAMS, S.C., HARIS, P.I., CHAPMAN, D.

and SIM, R.B.

The secondary structure of the von Willebrand factor type A domain in

10   factor B of human complement by Fourier transform infrared spectroscopy.

Its occurrence in collagen types VI, VII, XII and XIV, the integrins and

other proteins by averaged structure predictions.

J.MOL.BIOL. 238 104-119 (1994).

15   4. BORK, P. and ROHDE, K.

More von Willebrand factor type A domains? Sequence similarities with

malaria thrombospondin-related anonymous protein, dihydropyridine-

sensitive calcium channel and inter-alpha-trypsin inhibitor.

BIOCHEM.J. 279 908-910 (1991).

20

5. EDWARDS, Y.J.K. and PERKINS, S.J.

The protein fold of the von Willebrand factor type A domain is predicted

to be similar to the open twisted beta-sheet flanked by alpha-helices

found in human ras-p21.

25   FEBS LETT. 358 283-286 (1995).

6. LEE, J.O., RIEU, P., ARNAOUT, M.A. and LIDDINGTON, R.

Crystal structure of the A domain from the alpha subunit of integrin CR3

(CD11b/CD18).

30   CELL 80 631-638 (1995).

7. QU, A. and LEAHY, D.J.

Crystal structure of the I-domain from the CD11a/CD18 (LFA-1,

alpha L beta 2) integrin.

PROC.NATL.ACAD.SCI.USA 92 10277-10281 (1995).

The von Willebrand factor is a large multimeric glycoprotein found in blood
plasma. Mutant forms are involved in the aetiology of bleeding disorders
[1]. In von Willebrand factor, the type A domain (vWF) is the prototype for
a protein superfamily. The vWF domain is found in various plasma proteins:
complement factors B, C2, CR3 and CR4; the integrins (I-domains); collagen
types VI, VII, XII and XIV; and other extracellular proteins [2-4]. Proteins
that incorporate vWF domains participate in numerous biological events
(e.g., cell adhesion, migration, homing, pattern formation, and signal
transduction), involving interaction with a large array of ligands [2].
Secondary structure prediction from 75 aligned vWF sequences has revealed
a largely alternating sequence of alpha-helices and beta-strands [3]. Fold
recognition algorithms were used to score sequence compatibility with a
library of known structures: the vWF domain fold was predicted to be a
doubly-wound, open, twisted beta-sheet flanked by alpha-helices [5].
3D structures have been determined for the I-domains of integrins CD11b
(with bound magnesium) [6] and CD11a (with bound manganese) [7]. The domain
adopts a classic alpha/beta Rossmann fold and contains an unusual metal
ion coordination site at its surface. It has been suggested that this site
represents a general metal ion-dependent adhesion site (MIDAS) for binding
protein ligands [6]. The residues constituting the MIDAS motif in the CD11b
and CD11a I-domains are completely conserved, but the manner in which the
metal ion is coordinated differs slightly [7].

VWFADOMAIN is a 3-element fingerprint that provides a signature for the vWF
domain superfamily. The fingerprint was derived from an initial alignment
of 14 sequences. Motif 1 includes the first beta-strand and 3 conserved
residues involved in metal ion coordination in I-domains (Asp and 2 serines
in positions 8, 10 and 12, respectively); motif 2 spans strands beta-2 and
beta-2'; and motif 3 encodes beta-strand 3 and a conserved Asp (in position
7), which coordinates the metal ion [6,7]. Three iterations on OWL27.0 were
required to reach convergence, at which point a true set comprising 56
sequences was identified. Numerous partial matches were also found.

717. (WD40) WD domain, G-beta repeat

The ancient regulatory-protein family of WD-repeat proteins.

5   Neer EJ, Schmidt CJ, Nambudripad R, Smith TF;

Nature 1994;371:297-300.

Beta-transducin (G-beta) is one of the three subunits (alpha, beta, and gamma) of the guanine nucleotide-binding proteins (G proteins) which act as intermediaries in the transduction of signals generated by transmembrane

10   receptors [1]. The alpha subunit binds to and hydrolyzes GTP; the functions of the beta and gamma subunits are less clear but they seem to be required for the replacement of GDP by GTP as well as for membrane anchoring and receptor recognition.

15   In higher eukaryotes G-beta exists as a small multigene family of highly conserved proteins of about 340 amino acid residues. Structurally G-beta consists of eight tandem repeats of about 40 residues, each containing a central Trp-Asp motif (this type of repeat is sometimes called a WD-40 repeat). Such a repetitive segment has been shown [E1,2,3,4,5] to exist in a

20   number of other proteins listed below:

- Yeast STE4, a component of the pheromone response pathway. STE4 is a G-beta like protein that associates with GPA1 (G-alpha) and STE18 (G-gamma).
- Yeast MSI1, a negative regulator of RAS-mediated cAMP synthesis. MSI1 is

25   most probably also a G-beta protein.

- Human and chicken protein 12.3. The function of this protein is not known, but on the basis of its similarity to G-beta proteins, it may also function in signal transduction.

30   - Chlamydomonas reinhardtii gblp. This protein is most probably the homolog of vertebrate protein 12.3.

- Human LIS1, a neuronal protein involved in type-1 lissencephaly [E2].
- Mammalian coatomer beta' subunit (beta'-COP), a component of a cytosolic protein complex that reversibly associates with Golgi membranes to form

vesicles that mediate biosynthetic protein transport.

- Yeast CDC4, essential for initiation of DNA replication and separation of
  the spindle pole bodies to form the poles of the mitotic spindle.

5   - Yeast CDC20, a protein required for two microtubule-dependent processes:
  nuclear movements prior to anaphase and chromosome separation.

- Yeast MAK11, essential for cell growth and for the replication of M1
  double-stranded RNA.

- Yeast PRP4, a component of the U4/U6 small nuclear ribonucleoprotein with

10   a probable role in mRNA splicing.

- Yeast PWP1, a protein of unknown function.

- Yeast SKI8, a protein essential for controlling the propagation of double-
  stranded RNA.

- Yeast SOF1, a protein required for ribosomal RNA processing which

15   associates with U3 small nucleolar RNA.

- Yeast TUP1 (also known as AER2 or SFL2 or CYC9), a protein which has been
  implicated in dTMP uptake, catabolite repression, mating sterility, and
  many other phenotypes.

- Yeast YCR57c, an ORF of unknown function from chromosome III.

20   - Yeast YCR72c, an ORF of unknown function from chromosome III.

- Slime mold coronin, an actin-binding protein.

- Slime mold AAC3, a developmentally regulated protein of unknown function.

25   - Drosophila protein Groucho (formerly known as E(spl); 'enhancer of split'),
  a protein involved in neurogenesis and that seems to interact with the
  Notch and Delta proteins.

- Drosophila TAF-II-80, a protein that is tightly associated with TFIID.

30   The number of repeats in the above proteins varies between 5 (PRP4, TUP1, and
  Groucho) and 8 (G-beta, STE4, MSI1, AAC3, CDC4, PWP1, etc.). In G-beta and G-
  beta like proteins, the repeats span the entire length of the sequence, while
  in other proteins, they make up the N-terminal, the central or the C-terminal
  section.

A signature pattern can be developed from the central core of the domain (positions 9 to 23).

5 -Consensus pattern: [LIVMSTAC]-[LIVMFYWSTAGC]-[LIMSTAG]-[LIVMSTAGC]-x(2)-[DN]-

x(2)-[LIVMWSTAC]-x-[LIVMFSTAG]-W-[DEN]-[LIVMFSTAGCN]

[1] Gilman A.G.

10   Annu. Rev. Biochem. 56:615-649(1987).

[2] Duronio R.J., Gordon J.I., Boguski M.S.

Proteins 13:41-56(1992).

[3] van der Voorn L., Ploegh H.L.

FEBS Lett. 307:131-134(1992).

15 [4] Neer E.J., Schmidt C.J., Nambudripad R., Smith T.F.

Nature 371:297-300(1994).

[5] Smith T.F., Gaiatzes C.G., Saxena K., Neer E.J.

Biochemistry In Press(1998).

20

718. WHEP-TRS domain containing proteins

A conserved domain of 46 amino acids has been shown [1] to exist in a number of higher eukaryote aminoacyl-transfer RNA synthetases. This domain is present one to six times in the following enzymes:

25

- Mammalian multifunctional aminoacyl-tRNA synthetase. The domain is present three times in a region that separates the N-terminal glutamyl-tRNA synthetase domain from the C-terminal prolyl-tRNA synthetase domain.
- Drosophila multifunctional aminoacyl-tRNA synthetase. The domain is present

30   six times in the intercatalytic region.
- Mammalian tryptophanyl-tRNA synthetase. The domain is found at the N-terminal extremity.
- Mammalian, insect, nematode and plant glycyl-tRNA synthetase. The domain is found at the N-terminal extremity [2].

- Mammalian histidyl-tRNA synthetase. The domain is found at the N-terminal extremity.

This domain, which is called WHEP-TRS, could contain a central alpha-helical region and may play a role in the association of tRNA-synthetases into multienzyme complexes.

A signature pattern based on the first 29 positions of the WHEP-Domain has been developed.

-Consensus pattern: [QY]-G-[DNEA]-x-[LIV]-[KR]-x(2)-K-x(2)-[KRNG]-[AS]-x(4)-[LIV]-[DENK]-x(2)-[IV]-x(2)-L-x(3)-K

[1] Cerini C., Kerjan P., Astier M., Gratecos D., Mirande M., Semeriva M.
    EMBO J. 10:4267-4277(1991).
[2] Nada S., Chang P.K., Dignam J.D.
    J. Biol. Chem. 268:7660-7667(1993).

719. (Worm family 8) Putative membrane protein
Analysis of protein domain families in Caenorhabditis elegans.
Sonnhammer EL, Durbin R;
Genomics 1997;46:200-216.
This family called family 8 in [1], may be a transmembrane protein
The specific function of this protein is unknown.

720. Xylose isomerase
Xylose isomerase (EC 5.3.1.5) [1] is an enzyme found in microorganisms which catalyzes the interconversion of D-xylose to D-xylulose. It can also isomerize D-ribose to D-ribulose and D-glucose to D-fructose. Xylose isomerase seems to require magnesium for its activity, while cobalt is necessary to stabilize the tetrameric structure of the enzyme. A number of residues are conserved in all known xylose isomerases.

Xylose isomerase also exists in plants [2] where it is homodimeric and is manganese-dependent.

5   Two signatures patterns for xylose isomerase have been developed. The first one is derived from a stretch of five conserved amino acids that includes a glutamic acid residue known to be one of the four residues involved in the binding of the magnesium ion [3]; this pattern also includes a lysine residue which is involved in the catalytic activity. The second pattern is derived from a

10   conserved region in the N-terminal section of the enzyme that include an histidine residue which has been shown [4] to be involved in the catalytic mechanism of the enzyme.

-Consensus pattern: [LI]-E-P-K-P-x(2)-P

15   [E is a magnesium ligand]

[K is an active site residue]

-Consensus pattern: [FL]-H-D-x-D-[LIV]-x-[PD]-x-[GDE]

[H is an active site residue]

20   [1] Dauter Z., Dauter M., Hemker J., Witzel H., Wilson K.S.

FEBS Lett. 247:1-8(1989).

[2] Kristo P.A., Saarelainen R., Fagerstrom R., Aho S., Korhola M.

Eur. J. Biochem. 237:240-246(1996).

[3] Henrick K., Collyer C.A., Blow D.M.

25   J. Mol. Biol. 208:129-157(1989).

[4] Vangrysperre W., Ampe C., Kersters-Hilderson H., Tempst P.

Biochem. J. 263:195-199(1989).

30   721. XPG protein signatures. Xeroderma pigmentosum (XP) [1] is a human autosomal recessive disease, characterized by a high incidence of sunlight-induced skin cancer. People's skin cells with this condition are hypersensitive to ultraviolet light, due to defects in the incision step of DNA excision repair. There are a minimum of seven genetic complementation groups involved in this pathway: XP-A to XP-G. The defect in XP-G can

be corrected by a 133 Kd nuclear protein called XPG (or XPGC) [2].XPG belongs to a family of proteins [2,3,4,5,6] that are composed of twomain subsets: - Subset 1, to which belongs XPG, RAD2 from budding yeast and rad13 from fission yeast. RAD2 and XPG are single-stranded DNA endonucleases [7,8]. XPG makes the 3'incision in human DNA nucleotide

5   excision repair [9]. - Subset 2, to which belongs mouse and human FEN-1, rad2 from fission yeast, and RAD27 from budding yeast. FEN-1 is a structure-specific endonuclease. In addition to the proteins listed in the above groups, this family also includes: - Fission yeast exo1, a 5'->3' double-stranded DNA exonuclease that could act in a pathway that corrects mismatched base pairs. - Yeast EXO1 (DHS1), a protein with probably the same function as

10   exo1. - Yeast DIN7.Sequence alignment of this family of proteins reveals that similarities are largely confined to two regions. The first is located at the N-terminal extremity (N-region) and corresponds to the first 95 to 105 amino acids. The second region is internal (I-region) and found towards the C-terminus; it spans about 140 residues and contains a highly conserved core of 27 amino acids that includes a conserved pentapeptide (E-A-[DE]-A-[QS]).

15   It is possible that the conserved acidic residues are involved in the catalytic mechanism of DNA excision repair in XPG. The amino acids linking the N- and I-regions are not conserved; indeed, they are largely absent from proteins belonging to the second subset. Two signature patterns have been developed for these proteins. The first corresponds to the central part of the N-region, the second to part of the I-region and includes the putative catalytic core

20   pentapeptide

Consensus pattern: [VI]-[KRE]-P-x-[FYIL]-V-F-D-G-x(2)-[PIL]-x-[LVC]-K-
Consensus pattern: [GS]-[LIVM]-[PER]-[FYS]-[LIVM]-x-A-P-x-E-A-[DE]-[PAS]- [QS]-[CLM]-

25

[1] Tanaka K., Wood R.D. Trends Biochem. Sci. 19:83-86(1994).[2] Scherly D., Nouspikel T., Corlet J., Ucla C., Bairoch A., Clarkson S.G. Nature 363:182-185(1993).[3] Carr A.M., Sheldrick K.S., Murray J.M., Al-Harithy R., Watts F.Z., Lehmann A.R. Nucleic Acids Res. 21:1345-1349(1993).[4] Murray J.M., Tavassoli M., Al-Harithy R., Sheldrick K.S.,

30   Lehmann A.R., Carr A.M., Watts F.Z. Mol. Cell. Biol. 14:4878-4888(1994).[5] Harrington J.J., Lieber M.R. Genes Dev. 8:1344-1355(1994).[6] Szankasi P., Smith G.R. Science 267:1166-1169(1995).[7] Habraken Y., Sung P., Prakash L., Prakash S. Nature 366:365-368(1993).[8] O'Donovan A., Scherly D., Clarkson S.G., Wood R.D. J. Biol. Chem.

269:15965-15968(1994).[9] O'Donovan A., Davies A.A., Moggs J.G., West S.C., Wood R.D. Nature 371:432-435(1994).

5   722. Xanthine/uracil permeases family

The following transport proteins which are involved in the uptake of xanthine or uracil are evolutionary related [1]:

- Uric uric acid-xanthine permease (gene uapA) from Aspergillus nidulans.
10   - Purine permease (gene uapC) from Aspergillus nidulans.
- Xanthine permease from Bacillus subtilis (gene pbuX).
- Uracil permease from Escherichia coli (gene uraA) [2] and Bacillus (gene pyrP).
- Hypothetical protein ycdG from Escherichia coli.
15   - Hypothetical protein ygfO from Escherichia coli.
- Hypothetical protein ygfU from Escherichia coli.
- Hypothetical protein yicE from Escherichia coli.
- Hypothetical protein yunJ from Bacillus subtilis.
- Hypothetical protein yunK from Bacillus subtilis.
20

They are proteins of from 430 to 595 residues that seem to contain 12 transmembrane domains.
The best conserved region which corresponds with what seems to be the tenth transmembrane domain has been selected as a signature pattern.
25

-Consensus pattern: [LIVM]-P-x-[PASIF]-V-[LIVM]-G-G-x(4)-[LIVM]-[FY]-[GSA]-x-[LIVM]-x(3)-G
[1] Diallinas G., Gorfinkiel L., Arst G., Cecchetto G., Scazzocchio C.
J. Biol. Chem. 270:8610-8622(1995).
30   [2] Andersen P.S., Frees D., Fast R., Mygind B.
J. Bacteriol. 177:2008-2013(1995).

723. Hypothetical yabO/yceC/sfhB family

The following proteins, which seems to belong to a family of pseudouridine synthases (EC 4.2.1.70) [1] have been shown to share regions of similarities:

5
- Escherichia coli and Haemophilus influenzae ribosomal large subunit pseudouridine synthase A (gene rluA). It is responsible for synthesis of pseudouridine from uracil-746 IN 23S rRNA.

- Escherichia coli and Haemophilus influenzae ribosomal large subunit pseudouridine synthase C (gene rluC). It is responsible for synthesis of pseudouridine from uracil at positions 955, 2504 and 2580 in 23S rRNA.

10
- Escherichia coli protein and homologs in other bacteria large subunit pseudouridine synthase D (gene rluD).

- Yeast DRAP deaminase (gene RIB2).

- Escherichia coli hypothetical protein yqcB and HI1435, the corresponding Haemophilus influenzae protein.

15
- Haemophilus influenzae hypothetical protein HI0042.

- Aquifex aeolicus hypothetical protein AQ_1758.

- Bacillus subtilis hypothetical protein yhcT.

- Bacillus subtilis hypothetical protein yjbO.

- Bacillus subtilis hypothetical protein ylyB.

20
- Helicobacter pylori hypothetical protein HP0347.

- Helicobacter pylori hypothetical protein HP0745.

- Helicobacter pylori hypothetical protein HP0956.

- Mycoplasma genitalium hypothetical protein MG209.

- Mycoplasma genitalium hypothetical protein MG370.

25
- Synechocystis strain PCC 6803 hypothetical protein slr1592.

- Synechocystis strain PCC 6803 hypothetical protein slr1629.

- Yeast hypothetical protein YDL036c.

- Yeast hypothetical protein YGR169c.

- Fission yeast hypothetical protein SpAC18B11.02c.

30
- Caenorhabditis elegans hypothetical protein K07E8.7.

These are proteins of from 21 to 50 Kd which contain a number of conserved regions in their central section. They can be picked up in the database by the following highly conserved pattern.

-Consensus pattern: [LIVCA]-[NHYT]-R-[LI]-D-x(2)-T-[STA]-G-[LIVAGC]-[LIVMF](2)-[LIVMFGC]-[SGTACV]

5   [1] Conrad J., Sun D., Englund N., Ofengand J.
       J. Biol. Chem. 273:18562-18566(1998).

In addition, the following bacterial proteins, which seems to belong to a family of pseudouridine synthases (EC 4.2.1.70) [1] also have been shown to share regions of
10   similarities:

- Escherichia coli and Haemophilus influenzae 16S pseudouridylate 516 synthase (EC 4.2.1.70) (gene: rsuA). This enzyme is responsible for the formation of pseudouridine from uracil-516 in 16S ribosomal RNA.

15   - Escherichia coli hypothetical protein yciL and HI1199, the corresponding Haemophilus influenzae protein.

- Escherichia coli hypothetical protein yjbC.

- Escherichia coli hypothetical protein ymfC and HI0694, the corresponding Haemophilus influenzae protein.

20   - Aquifex aeolicus hypothetical protein AQ_554.

- Aquifex aeolicus hypothetical protein AQ_1464.

- Bacillus subtilis hypothetical protein ypuL.

- Bacillus subtilis hypothetical protein ytzF.

- Borrelia burgdorferi hypothetical protein BB0129.

25   - Helicobacter pylori hypothetical protein HP1459.

- Synechocystis strain PCC 6803 hypothetical protein slr0361.

- Synechocystis strain PCC 6803 hypothetical protein slr0612.

These are proteins of from 25 to 40 Kd which contain a number of conserved
30   regions in their central section. They can be picked up in the database by the following highly conserved pattern.

-Consensus pattern: G-R-L-D-x(2)-[STA]-x-G-[LIVFA]-[LIVMF](3)-[ST]-[DNST]

561

[1] Wrzesinski J., Bakin A., Nurse K., Lane B.G., Ofengand J.
Biochemistry 34:8904-8913(1995).

5   724. Zinc finger present in dystrophin, CBP/p300

ZZ in dystrophin binds calmodulin

Putative zinc finger; binding not yet shown.

10   725. Zinc carboxypeptidase

There are a number of different types of zinc-dependent carboxypeptidases (EC
3.4.17.-) [1,2]. All these enzymes seem to be structurally and functionally
related. The enzymes that belong to this family are listed below.

15   - Carboxypeptidase A1 (EC 3.4.17.1), a pancreatic digestive enzyme that can
removes all C-terminal amino acids with the exception of Arg, Lys and Pro.

- Carboxypeptidase A2 (EC 3.4.17.15), a pancreatic digestive enzyme with a
specificity similar to that of carboxypeptidase A1, but with a preference
for bulkier C-terminal residues.

20   - Carboxypeptidase B (EC 3.4.17.2), also a pancreatic digestive enzyme, but
that preferentially removes C-terminal Arg and Lys.

- Carboxypeptidase N (EC 3.4.17.3) (also known as arginine carboxypeptidase),
a plasma enzyme which protects the body from potent vasoactive and
inflammatory peptides containing C-terminal Arg or Lys (such as kinins or

25   anaphylatoxins) which are released into the circulation.

- Carboxypeptidase H (EC 3.4.17.10) (also known as enkephalin convertase or
carboxypeptidase E), an enzyme located in secretory granules of pancreatic
islets, adrenal gland, pituitary and brain. This enzyme removes residual C-
terminal Arg or Lys remaining after initial endoprotease cleavage during

30   prohormone processing.

- Carboxypeptidase M (EC 3.4.17.12), a membrane bound Arg and Lys specific
enzyme.

It is ideally situated to act on peptide hormones at local tissue sites
where it could control their activity before or after interaction with

specific plasma membrane receptors.

- Mast cell carboxypeptidase (EC 3.4.17.1), an enzyme with a specificity
to carboxypeptidase A, but found in the secretory granules of mast cells.

- Streptomyces griseus carboxypeptidase (Cpase SG) (EC 3.4.17.-) [3], which
5   combines the specificities of mammalian carboxypeptidases A and B.

- Thermoactinomyces vulgaris carboxypeptidase T (EC 3.4.17.18) (CPT) [4],
which also combines the specificities of carboxypeptidases A and B.

- AEBP1 [5], a transcriptional repressor active in preadipocytes. AEBP1 seems
to regulate transcription by cleavage of other transcriptional proteins.

10   - Yeast hypothetical protein YHR132c.

All of these enzymes bind an atom of zinc. Three conserved residues are
implicated in the binding of the zinc atom: two histidines and a glutamic acid
Two signature patterns which contain these three zinc-ligands have been derived.

15

-Consensus pattern: [PK]-x-[LIVMFY]-x-[LIVMFY]-x(4)-H-[STAG]-x-E-x-[LIVM]-
[STAG]-x(6)-[LIVMFYTA]
[H and E are zinc ligands]
-Consensus pattern: H-[STAG]-x(3)-[LIVME]-x(2)-[LIVMFYW]-P-[FYW]
20   [H is a zinc ligand]

[1] Tan F., Chan S.J., Steiner D.F., Schilling J.W., Skidgel R.A.
   J. Biol. Chem. 264:13165-13170(1989).
[2] Reynolds D.S., Stevens R.L., Gurley D.S., Lane W.S., Austen K.F.,
25   Serafin W.E.
   J. Biol. Chem. 264:20094-20099(1989).
[3] Narahashi Y.
   J. Biochem. 107:879-886(1990).
[4] Teplyakov A., Polyakov K., Obmolova G., Strokopytov B., Kuranova I.,
30   Osterman A.L., Grishin N.V., Smulevitch S.V., Zagnitko O.P.,
   Galperina O.V., Matz M.V., Stepanov V.M.
   Eur. J. Biochem. 208:281-288(1992).
[5] He G.-P., Muise A., Li A.W., Ro H.-S.
   Nature 378:92-96(1995).

563

[6] Hourdou M.-L., Guinand M., Vacheron M.J., Michel G., Denoroy L.,

Duez C.M., Englebert S., Joris B., Weber G., Ghuysen J.-M.

Biochem. J. 292:563-570(1993).

[7] Rawlings N.D., Barrett A.J.

5   Meth. Enzymol. 248:183-228(1995).

726. Zinc finger, C2H2 type

The C2H2 zinc finger is the classical zinc finger domain.

10   The two conserved cysteines and histidines co-ordinate a

zinc ion. The following pattern describes the zinc finger.

#-X-C-X(1-5)-C-X3-#-X5-#-X2-H-X(3-6)-[H/C]

Where X can be any amino acid, and numbers in brackets

indicate the number of residues. The positions marked # are

15   those that are important for the stable fold of the zinc

finger. The final position can be either his or cys.

The C2H2 zinc finger is composed of two short beta strands

followed by an alpha helix. The amino terminal part of the

helix binds the major groove in DNA binding zinc fingers.

20

'Zinc finger' domains [1-5] are nucleic acid-binding protein structures first

identified in the Xenopus transcription factor TFIIIA. These domains have

since been found in numerous nucleic acid-binding proteins. A zinc finger

domain is composed of 25 to 30 amino-acid residues. There are two cysteine or

25   histidine residues at both extremities of the domain, which are involved in

the tetrahedral coordination of a zinc atom. It has been proposed that such a

domain interacts with about five nucleotides. A schematic representation of a

zinc finger domain is shown below:

30   X X

      X X

      X X

      X X

      X X

```
   x  x

   C  H

   x  \  /  x

   x  Zn  x

   x  /  \  x

   C  H

   x x x x x  x x x x x
```

Many classes of zinc fingers are characterized according to the number and positions of the histidine and cysteine residues involved in the zinc atom coordination. In the first class to be characterized, called C2H2, the first pair of zinc coordinating residues are cysteines, while the second pair are histidines. A number of experimental reports have demonstrated the zinc-dependent DNA or RNA binding property of some members of this class.

Some of the proteins known to include C2H2-type zinc fingers are listed below. The number of zinc finger regions found in each of these proteins are indicated between brackets; a '+' symbol indicates that only partial sequence data is available and that additional finger domains may be present.

- Saccharomyces cerevisiae: ACE2 (3), ADR1 (2), AZF1 (4), FZF1 (5), MIG1 (2), MSN2 (2), MSN4 (2), RGM1 (2), RIM1 (3), RME1 (3), SFP1 (2), SSL1 (1), STP1 (3), SWI5 (3), VAC1 (1) and ZMS1 (2).
- Emericella nidulans: brlA (2), creA (2).
- Drosophila: AEF-1 (4), Cf2 (7), ci-D (5), Disconnected (2), Escargot (5), Glass (5), Hunchback (6), Kruppel (5), Kruppel-H (4+), Odd-skipped (4), Odd-paired (4), Pep (3), Snail (5), Spalt-major (7), Serependity locus beta (6), delta (7), h-1 (8), Suppressor of hairy wing su(Hw) (12), Suppressor of variegation suvar(3)7 (5), Teashirt (3) and Tramtrack (2).
- Xenopus: transcription factor TFIIIA (9), p43 from RNP particle (9), Xfin (37 !!), Xsna (5), gastrula XlcGF5.1 to XlcGF71.1 (from 4+ to 11+), Oocyte XlcOF2 to XlcOF22 (from 7 to 12).
- Mammalian: basonuclin (6), BCL-6/LAZ-3 (6), erythroid k)rueppel-like transcription factor (3), transcription factors Sp1 (3), Sp2 (3), Sp3 (3)

and Sp(4) 3, transcriptional repressor YY1 (4), Wilms' tumor protein (4), EGR1/Krox24 (3), EGR2/Krox20 (3), EGR3/Pilot (3), EGR4/AT133 (4), Evi-1 (10), GLI1 (5), GLI2 (4+), GLI3 (3+), HIV-EP1/ZNF40 (4), HIV-EP2 (2), KR1 (9+), KR2 (9), KR3 (15+), KR4 (14+), KR5 (11+), HF.12 (6+), REX-1 (4), ZfX (13), ZfY (13), Zfp-35 (18), ZNF7 (15), ZNF8 (7), ZNF35 (10), ZNF42/MZF-1 (13), ZNF43 (22), ZNF46/Kup (2), ZNF76 (7), ZNF91 (36), ZNF133 (3).

In addition to the conserved zinc ligand residues it has been shown [6] that a number of other positions are also important for the structural integrity of the C2H2 zinc fingers. The best conserved position is found four residues after the second cysteine; it is generally an aromatic or aliphatic residue.

-Consensus pattern: C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

[The two C's and two H's are zinc ligands]

[1] Klug A., Rhodes D.
   Trends Biochem. Sci. 12:464-469(1987).
[2] Evans R.M., Hollenberg S.M.
   Cell 52:1-3(1988).
[3] Payre F., Vincent A.
   FEBS Lett. 234:245-250(1988).
[4] Miller J., McLachlan A.D., Klug A.
   EMBO J. 4:1609-1614(1985).
[5] Berg J.M.
   Proc. Natl. Acad. Sci. U.S.A. 85:99-102(1988).
[6] Rosenfeld R., Margalit H.
   J. Biomol. Struct. Dyn. 11:557-570(1993).

727. Zinc finger, C3HC4 type (RING finger)

A number of eukaryotic and viral proteins contain a conserved cysteine-rich domain of 40 to 60 residues (called C3HC4 zinc-finger or 'RING' finger) [1] that binds two atoms of zinc, and is probably involved in mediating protein-protein interactions. The 3D structure of the zinc ligation system is unique

to the RING domain and is refered to as the "cross-brace" motif. The spacing of the cysteines in such a domain is C-x(2)-C-x(9 to 39)-C-x(1 to 3)-H-x(2 to 3)-C-x(2)-C-x(4 to 48)-C-x(2)-C.

5   Proteins currently known to include the C3HC4 domain are listed below (references are only provided for recently determined sequences).

- Mammalian V(D)J recombination activating protein (gene RAG1). RAG1 activates the rearrangement of immunoglobulin and T-cell receptor genes.

10   - Mouse rpt-1. Rpt-1 is a trans-acting factor that regulates gene expression directed by the promoter region of the interleukin-2 receptor alpha chain or the LTR promoter region of HIV-1.

- Human rfp. Rfp is a developmentally regulated protein that may function in male germ cell development. Recombination of the N-terminal section of rfp

15   with a protein tyrosine kinase produces the ret transforming protein.

- Human 52 Kd Ro/SS-A protein. A protein of unknown function from the Ro/SS-A ribonucleoprotein complex. Sera from patients with systemic lupus erythematosus or primary Sjogren's syndrome often contain antibodies that react with the Ro proteins.

20   - Human histocompatibility locus protein RING1.

- Human PML, a probable transcription factor. Chromosomal translocation of PML with retinoic receptor alpha creates a fusion protein which is the cause of acute promyelocytic leukemia (APL).

- Mammalian breast cancer type 1 susceptibility protein (BRCA1) [E1].

25   - Mammalian cbl proto-oncogene.

- Mammalian bmi-1 proto-oncogene.

- Vertebrate CDK-activating kinase (CAK) assembly factor MAT1, a protein that stabilizes the complex between the CDK7 kinase and cyclin H (MAT1 stands for 'Menage A Trois').

30   - Mammalian mel-18 protein. Mel-18 which is expressed in a variety of tumor cells is a transcriptional repressor that recognizes and bind a specific DNA sequence.

- Mammalian peroxisome assembly factor-1 (PAF-1) (PMP35), which is somewhat involved in the biogenesis of peroxisomes. In humans, defects in PAF-1 are

responsible for a form of Zellweger syndrome, an autosomal recessive disorder associated with peroxisomal deficiencies.

- Human MAT1 protein, which interacts with the CDK7-cyclin H complex.

- Human RING1 protein.

5   - Xenopus XNF7 protein, a probable transcription factor.

- Trypanosoma protein ESAG-8 (T-LR), which may be involved in the postranscriptional regulation of genes in VSG expression sites or may interact with adenylate cyclase to regulate its activity.

- Drosophila proteins Posterior Sex Combs (Psc) and Suppressor two of zeste

10   (Su(z)2). The two proteins belong to the Polycomb group of genes needed to maintain the segment-specific repression of homeotic selector genes.

- Drosophila protein male-specific msl-2, a DNA-binding protein which is involved in X chromosome dosage compensation (the elevation of transcription of the male single X chromosome).

15   - Arabidopsis thaliana protein COP1 which is involved in the regulation of photomorphogenesis.

- Fungal DNA repair proteins RAD5, RAD16, RAD18 and rad8.

- Herpesviruses trans-acting transcriptional protein ICP0/IE110. This protein which has been characterized in many different herpesviruses is a trans-

20   activator and/or -repressor of the expression of many viral and cellular promoters.

- Baculoviruses protein CG30.

- Baculoviruses major immediate early protein (PE-38).

- Baculoviruses immediate-early regulatory protein IE-N/IE-2.

25   - Caenorhabditis elegans hypothetical proteins F54G8.4, R05D3.4 and T02C1.1.

- Yeast hypothetical proteins YER116c and YKR017c.

The central region of the domain was selected as a signature pattern for the C3HC4 finger.

30

-Consensus pattern: C-x-H-x-[LIVMFY]-C-x(2)-C-[LIVMYA]

[1] Borden K.L.B., Freemont P.S.

Curr. Opin. Struct. Biol. 6:395-401(1996).

728. Zinc finger C-x8-C-x5-C-x3-H type (and similar).

5

729. Zinc finger, CCHC class

A family of CCHC zinc fingers, mostly from retroviral gag

proteins (nucleocapsid). Prototype structure is from HIV.

Also contains members involved in eukaryotic gene

10   regulation, such as C. elegans GLH-1.

Structure is an 18-residue zinc finger; no examples of indels

in the alignment.

15   730. Zn-finger in Ran binding protein and others.

731. AN1-like Zinc finger

20   Zinc finger at the C-terminus of An1 Swiss:Q91889, a ubiquitin-like protein in Xenopus

laevis. The following pattern describes the zinc finger. C-X2-C-X(9-12)-C-X(1-2)-C-X4-C-

X2-H-X5-H-X-C Where X can be any amino acid, and numbers in brackets indicate the

number of residues.

25   [1] Linnen JM, Bailey CP, Weeks DL; Gene 1993;128:181-188.

732. 14-3-3 proteins

Structure of a 14-3-3 protein and implications for coordination of multiple

30   signalling pathways.

Xiao B, Smerdon SJ, Jones DH, Dodson GG, Soneji Y, Aitken A, Gamblin SJ;

Nature 1995;376:188-191.

Crystal structure of the zeta isoform of the 14-3-3 protein.

Liu D, Bienkowska J, Petosa C, Collier RJ, Fu H, Liddington R;

Nature 1995;376:191-194.

Interaction of 14-3-3 with signaling proteins is mediated by the recognition of phosphoserine.

5   Muslin AJ, Tanner JW, Allen PM, Shaw AS;
Cell 1996;84:889-897.

The 14-3-3 protein binds its target proteins with a common site located towards the C-terminus.

10   Ichimura T, Ito M, Itagaki C, Takahashi M, Horigome T, Omata S, Ohno S, Isobe T
FEBS Lett 1997;413:273-276.

Molecular evolution of the 14-3-3 protein family.

15   Wang W, Shakes DC
J Mol Evol 1996;43:384-398.
Function of 14-3-3 proteins.
Jin DY, Lyu MS, Kozak CA, Jeang KT
Nature 1996;382:308-308.

20

The 14-3-3 proteins [1,2,3] are a family of closely related acidic homodimeric proteins of about 30 Kd which were first identified as being very abundant in mammalian brain tissues and located preferentially in neurons. The 14-3-3 proteins seem to have multiple biological activities and play a key role in

25   signal transduction pathways and the cell cycle. They interacts with kinases such as PKC or Raf-1; they seem to also function as protein-kinase dependent activators of tyrosine and tryptophan hydroxylases and in plants they are associated with a complex that binds to the G-box promoter elements.

30   The 14-3-3 family of proteins are ubiquitously found in all eukaryotic species studied and have been sequenced in fungi (yeast BMH1 and BMH2, fission yeast rad24 and rad25), plants, Drosophila, and vertebrates. The sequences of the 14-3-3 proteins are extremely well conserved. Two highly conserved regions have been selected as signature patterns: the first is a peptide of 11 residues

located in the N-terminal section; the second, a 20 amino acid region located in the C-terminal section.

-Consensus pattern: R-N-L-[LIV]-S-[VG]-[GA]-Y-[KN]-N-[IVA]

5   -Consensus pattern: Y-K-[DE]-S-T-L-I-[IM]-Q-L-[LF]-[RHC]-D-N-[LF]-T-[LS]-W-[TAN]-[SAD]

[1] Aitken A.

Trends Biochem. Sci. 20:95-97(1995).

10   [2] Morrison D.

Science 266:56-57(1994).

[3] Xiao B., Smerdon S.J., Jones D.H., Dodson G.G., Soneji Y., Aitken A.,

Gamblin S.J.

Nature 376:188-191(1995).

15

733. D-isomer specific 2-hydroxyacid dehydrogenases (2 Hacid DH)

This Pfam covers the Formate dehydrogenase, D-glycerate dehydrogenase and D-lactate dehydrogenase families in SCOP. A number of NAD-dependent 2-hydroxyacid dehydrogenases which seem to be specific for the D-isomer of their substrate have been shown [1,2,3,4] to be functionally and structurally related. These enzymes are listed below.

-    D-lactate dehydrogenase (EC 1.1.1.28), a bacterial enzyme which catalyzes the reduction of D-lactate to pyruvate.

-    D-glycerate dehydrogenase (EC 1.1.1.29) (NADH-dependent hydroxypyruvate reductase), a plant leaf peroxisomal enzyme that catalyzes the reduction of hydroxypyruvate to glycerate. This reaction is part of the glycolate pathway of photorespiration.

-    D-glycerate dehydrogenase from the bacteria Hyphomicrobium methylovorum and Methylobacterium extorquens.

-    3-phosphoglycerate dehydrogenase (EC 1.1.1.95), a bacterial enzyme that catalyzes the oxidation of D-3-phosphoglycerate to 3-phosphohydroxypyruvate. This reaction is the first committed step in the 'phosphorylated' pathway of serine biosynthesis.

571

- Erythronate-4-phosphate dehydrogenase (EC 1.1.1.-) (gene pdxB), a bacterial enzyme involved in the biosynthesis of pyridoxine (vitamin B6).

- D-2-hydroxyisocaproate dehydrogenase (EC 1.1.1.-) (D-hicDH), a bacterial enzyme that catalyzes the reversible and stereospecific interconversion between 2-ketocarboxylic acids and D-2-hydroxy-carboxylic acids.

- Formate dehydrogenase (EC 1.2.1.2) (FDH) from the bacteria Pseudomonas sp. 101 and various fungi [5].

- Vancomycin resistance protein vanH from Enterococcus faecium; this protein is a D-specific alpha-keto acid dehydrogenase involved in the formation of a peptidoglycan which does not terminate by D-alanine thus preventing vancomycin binding.

- Escherichia coli hypothetical protein ycdW.

- Escherichia coli hypothetical protein yiaE.

- Haemophilus influenzae hypothetical protein HI1556.

- Yeast hypothetical protein YER081w.

- Yeast hypothetical protein YIL074w.

All these enzymes have similar enzymatic activities and are structurally related. Three of the most conserved regions of these proteins have been selected to develop patterns. The first pattern is based on a glycine-rich region located in the central section of these enzymes; this region probably corresponds to the NAD-binding domain. The two other patterns contain a number of conserved charged residues, some of which may play a role in the catalytic mechanism.

-Consensus pattern: [LIVMA]-[AG]-[IVT]-[LIVMFY]-[AG]-x-G-[NHKRQGSAC]-[LIV]-G-x(13,14)-[LIVfMT]-x(2)-[FYwCTH]-[DNSTK]

-Consensus pattern: [LIVMFYWA]-[LIVFYWC]-x(2)-[SAC]-[DNQHR]-[IVFA]-[LIVF]-x-[LIVF]-[HNI]-x-P-x(4)-[STN]-x(2)-[LIVMF]-x-[GSDN]

-Consensus pattern: [LMFATC]-[KPQ]-x-[GSTDN]-x-[LIVMFYWR]-[LIVMFYW](2)-N-x-[STAGC]-R-[GP]-x-[LIVH]-[LIVMC]-[DNV]

[1] Grant G.A. Biochem. Biophys. Res. Commun. 165:1371-1374(1989).

[2] Kochhar S., Hunziker P., Leong-Morgenthaler P.M., Hottinger H. Biochem. Biophys. Res. Commun. 184:60-66(1992).

[3] Ohta T., Taguchi H. J. Biol. Chem. 266:12588-12594(1991).

[4] Goldberg J.D., Yoshida T., Brick P. J. Mol. Biol. 236:1123-1140(1994).

[5] Popov V.O., Lamzin V.S. Biochem. J. 301:625-643(1994).

5   734. 2-oxo acid dehydrogenases acyltransferase (catalytic domain)

Refined crystal structure of the catalytic domain of dihysrolipoyl

transacetylase (E2P) from azotobacter vineelandii at 2.6 angstroms

resolution.

Mattevi A, Obmolova G, Kalk KH, Westphal AH, De Kok A, Hol WG;

10   J Mol Biol 1993;230:1183-1199.

These proteins contain one to three copies of a lipoyl binding domain

followed by the catalytic domain.

15   735. 3-beta hydroxysteriod dehydrogenase/isomerase family

Structure and tissue-specific expression of 3

beta-hydroxysteroid dehydrogenase/5-ene-4-ene isomerase

genes in human and rat classical and peripheral

steroidogenic tissues.

20   Labrie F, Simard J, Luu-The V, Pelletier G, Belanger A,

Lachance Y, Zhao HF, Labrie C, Breton N, de Launoit Y, et al

J Steroid Biochem Mol Biol 1992;41:421-435.

The enzyme 3 beta-hydroxysteroid dehydrogenase/5-ene-4-ene

isomerase (3 beta-HSD) catalyzes the oxidation and isomerization

25   of 5-ene-3 beta-hydroxypregnene and 5-ene-hydroxyandrostene

steroid precursors into the corresponding 4-ene-ketosteroids necessary

for the formation of all classes of steroid hormones.

30   736. 3-hydroxyacyl-CoA dehydrogenase

This family also includes lambda crystallin.

Structure of L-3-hydroxyacyl-coenzyme A dehydrogenase:

preliminary chain tracing at 2.8-A resolution.

Birktoft JJ, Holden HM, Hamlin R, Xuong NH, Banaszak LJ;

Proc Natl Acad Sci U S A 1987;84:8262-8266.

3-hydroxyacyl-CoA dehydrogenase (EC 1.1.1.35) (HCDH) [1] is an enzyme involved in fatty acid metabolism, it catalyzes the reduction of 3-hydroxyacyl-CoA to

5   3-oxoacyl-CoA. Most eukaryotic cells have 2 fatty-acid beta-oxidation systems, one located in mitochondria and the other in peroxisomes. In peroxisomes 3-hydroxyacyl-CoA dehydrogenase forms, with enoyl-CoA hydratase (ECH) and 3,2-trans-enoyl-CoA isomerase (ECI) a multifunctional enzyme where the N-terminal domain bears the hydratase/isomerase activities and the C-terminal

10   domain the dehydrogenase activity. There are two mitochondrial enzymes: one which is monofunctional and the other which is, like its peroxisomal counterpart, multifunctional.

In Escherichia coli (gene fadB) and Pseudomonas fragi (gene faoA) HCDH is part

15   of a multifunctional enzyme which also contains an ECH/ECI domain as well as a 3-hydroxybutyryl-CoA epimerase domain [2].

The other proteins structurally related to HCDH are:

20   - Bacterial 3-hydroxybutyryl-CoA dehydrogenase (EC 1.1.1.157) which reduces 3-hydroxybutanoyl-CoA to acetoacetyl-CoA [3].
     - Eye lens protein lambda-crystallin [4], which is specific to lagomorphes (such as rabbit).

25   There are two major region of similarities in the sequences of proteins of the HCDH family, the first one located in the N-terminal, corresponds to the NAD-binding site, the second one is located in the center of the sequence. A signature pattern has been derived from this central region.

30   -Consensus pattern: [DNE]-x(2)-[GA]-F-[LIVMFY]-x-[NT]-R-x(3)-[PA]-[LIVMFY](2)-x(5)-[LIVMFYCT]-[LIVMFY]-x(2)-[GV]

[1] Birktoff J.J., Holden H.M., Hamlin R., Xuong N.-H., Banaszak L.J.
     Proc. Natl. Acad. Sci. U.S.A. 84:8262-8266(1987).

[2] Nakahigashi K., Inokuchi H.

Nucleic Acids Res. 18:4937-4937(1990).

[3] Mullany P., Clayton C.L., Pallen M.J., Slone R., Al-Saleh A.,

Tabaqchali S.

FEMS Microbiol. Lett. 124:61-67(1994).

[4] Mulders J.W.M., Hendriks W., Blankesteijn W.M., Bloemendal H.,

de Jong W.W.

J. Biol. Chem. 263:15462-15466(1988).

737. 60s Acidic ribosomal protein

Proteins P1, P2, and P0, components of the eukaryotic

ribosome stalk. New structural and functional aspects.

Remacha M, Jimenez-Diaz A, Santos C, Briones E, Zambrano R,

Rodriguez Gabriel MA, Guarinos E, Ballesta JP;

Biochem Cell Biol 1995;73:959-968.

This family includes archaebacterial L12, eukaryotic P0, P1 and P2.

738. 6-phosphogluconate dehydrogenases

6-phosphogluconate dehydrogenase (EC 1.1.1.44) (6PGD) catalyzes the third step
in the hexose monophosphate shunt, the decarboxylating reduction of
6-phosphogluconate in to ribulose 5-phosphate.

Prokaryotic and eukaryotic 6PGD are proteins of about 470 amino acids whose

sequence are highly conserved [1]. A region which has been shown [2], from studies

of the sheep 6PGD tertiary structure, to be involved in the binding of 6-phosphogluconate

has been selected as a signature pattern.

-Consensus pattern: [LIVM]-x-D-x(2)-[GA]-[NQS]-K-G-T-G-x-W

[1] Reizer A., Deutscher J., Saier M.H. Jr., Reizer J.

Mol. Microbiol. 5:1081-1089(1991).

[2] Adams M.J., Archibald I.G., Bugg C.E., Carne A., Gover S.,

Helliwell J.R., Pickersgill R.W., White S.W.

EMBO J. 2:1009-1014(1983).

5   739. (7tm 1) G-protein coupled receptors [1 to 4,E1,E2] (also called R7G) are an extensive

group of hormones, neurotransmitters, odorants and light receptors which

transduce extracellular signals by interaction with guanine nucleotide-

binding (G) proteins. The receptors that are currently known to belong to this

family are listed below.

10

- 5-hydroxytryptamine (serotonin) 1A to 1F, 2A to 2C, 4, 5A, 5B, 6 and 7 [5].

- Acetylcholine, muscarinic-type, M1 to M5.

- Adenosine A1, A2A, A2B and A3 [6].

- Adrenergic alpha-1A to -1C; alpha-2A to -2D; beta-1 to -3 [7].

15   - Angiotensin II types I and II.

- Bombesin subtypes 3 and 4.

- Bradykinin B1 and B2.

- c3a and C5a anaphylatoxin.

- Cannabinoid CB1 and CB2.

20   - Chemokines C-C CC-CKR-1 to CC-CKR-8.

- Chemokines C-X-C CXC-CKR-1 to CXC-CKR-4.

- Cholecystokinin-A and cholecystokinin-B/gastrin.

- Dopamine D1 to D5 [8].

- Endothelin ET-a and ET-b [9].

25   - fMet-Leu-Phe (fMLP) (N-formyl peptide).

- Follicle stimulating hormone (FSH-R) [10].

- Galanin.

- Gastrin-releasing peptide (GRP-R).

- Gonadotropin-releasing hormone (GNRH-R).

30   - Histamine H1 and H2 (gastric receptor I).

- Lutropin-choriogonadotropic hormone (LSH-R) [10].

- Melanocortin MC1R to MC5R.

- Melatonin.

- Neuromedin B (NMB-R).

- Neuromedin K (NK-3R).

- Neuropeptide Y types 1 to 6.

- Neurotensin (NT-R).

- Octopamine (tyramine), from insects.

5 - Odorants [11].

- Opioids delta-, kappa- and mu-types [12].

- Oxytocin (OT-R).

- Platelet activating factor (PAF-R).

- Prostacyclin.

10 - Prostaglandin D2.

- Prostaglandin E2, EP1 to EP4 subtypes.

- Prostaglandin F2.

- Purinoreceptors (ATP) [13].

- Somatostatin types 1 to 5.

15 - Substance-K (NK-2R).

- Substance-P (NK-1R).

- Thrombin.

- Thromboxane A2.

- Thyrotropin (TSH-R) [10].

20 - Thyrotropin releasing factor (TRH-R).

- Vasopressin V1a, V1b and V2.

- Visual pigments (opsins and rhodopsin) [14].

- Proto-oncogene mas.

- A number of orphan receptors (whose ligand is not known) from mammals and

25 birds.

- Caenorhabditis elegans putative   receptors C06G4.5, C38C10.1, C43C3.2, T27D1.3 and ZC84.4.

- Three putative receptors encoded in the genome of cytomegalovirus: US27, US28, and UL33.

30 - ECRF3, a putative receptor encoded in the genome of herpesvirus saimiri.

The structure of all these receptors is thought to be identical. They have seven hydrophobic regions, each of which most probably spans the membrane. The N-terminus is located on the extracellular side of the membrane and is

often glycosylated, while the C-terminus is cytoplasmic and generally phosphorylated. Three extracellular loops alternate with three intracellular loops to link the seven transmembrane regions. Most, but not all of these receptors, lack a signal peptide. The most conserved parts of these proteins

5   are the transmembrane regions and the first two cytoplasmic loops. A conserved acidic-Arg-aromatic triplet is present in the N-terminal extremity of the second cytoplasmic loop [15] and could be implicated in the interaction with G proteins.

10   To detect this widespread family of proteins, a pattern that contains the conserved triplet and that also spans the major part of the third transmembrane helix has been developed.

-Consensus pattern: [GSTALIVMFYWC]-[GSTANCPDE]-{EDPKRH}-x(2)-
15   [LIVMNQGA]-x(2)-
        [LIVMFT]-[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSH]-x(2)-
        [LIVM]

[1] Strosberg A.D.
20   Eur. J. Biochem. 196:1-10(1991).
     [2] Kerlavage A.R.
        Curr. Opin. Struct. Biol. 1:394-401(1991).
     [3] Probst W.C., Snyder L.A., Schuster D.I., Brosius J., Sealfon S.C.
        DNA Cell Biol. 11:1-20(1992).
25   [4] Savarese T.M., Fraser C.M.
        Biochem. J. 283:1-9(1992).
     [5] Branchek T.
        Curr. Biol. 3:315-317(1993).
     [6] Stiles G.L.
30   J. Biol. Chem. 267:6451-6454(1992).
     [7] Friell T., Kobilka B.K., Lefkowitz R.J., Caron M.G.
        Trends Neurosci. 11:321-324(1988).
     [8] Stevens C.F.
        Curr. Biol. 1:20-22(1991).

[9] Sakurai T., Yanagisawa M., Masaki T.

Trends Pharmacol. Sci. 13:103-107(1992).

[10] Salesse R., Remy J.J., Levin J.M., Jallal B., Garnier J.

Biochimie 73:109-120(1991).

5   [11] Lancet D., Ben-Arie N.

Curr. Biol. 3:668-674(1993).

[12] Uhl G.R., Childers S., Pasternak G.

Trends Neurosci. 17:89-93(1994).

[13] Barnard E.A., Burnstock G., Webb T.E.

10   Trends Pharmacol. Sci. 15:67-70(1994).

[14] Applebury M.L., Hargrave P.A.

Vision Res. 26:1881-1895(1986).

[15] Attwood T.K., Eliopoulos E.E., Findlay J.B.C.

Gene 98:153-159(1991).

15

(7tm 1) Visual pigments (opsins) retinal binding site

Visual pigments [1,2] are the light-absorbing molecules that mediate vision. They consist of an apoprotein, opsin, covalently linked to the chromophore cis-retinal. Vision is effected through the absorption of a photon by cis-

20   retinal which is isomerized to trans-retinal. This isomerization leads to a change of conformation of the protein. Opsins are integral membrane proteins with seven transmembrane regions that belong to family 1 of G-protein coupled receptors.

25   In vertebrates four different pigments are generally found. Rod cells, which mediate vision in dim light, contain the pigment rhodopsin. Cone cells, which function in bright light, are responsible for color vision and contain three or more color pigments (for example, in mammals: red, blue and green).

30   In Drosophila, the eye is composed of 800 facets or ommatidia. Each ommatidium contains eight photoreceptor cells (R1-R8): the R1 to R6 cells are outer cells, R7 and R8 inner cells. Each of the three types of cells (R1-R6, R7 and R8) expresses a specific opsin.

Proteins evolutionary related to opsins include squid retinochrome, also known as retinal photoisomerase, which converts various isomers of retinal into 11-cis retinal and mammalian retinal pigment epithelium (RPE) RGR [3], a protein that may also act in retinal isomerization.

5

The attachment site for retinal in the above proteins is a conserved lysine residue in the middle of the seventh transmembrane helix. The pattern that had been developed includes this residue.

10 -Consensus pattern: [LIVMWAC]-[PGC]-x(3)-[SAC]-K-[STALIMR]-[GSACPNV]-[STACP]-

x(2)-[DENF]-[AP]-x(2)-[IY]

[K is the retinal binding site]

15 [1] Applebury M.L., Hargrave P.A.

Vision Res. 26:1881-1895(1986).

[2] Fryxell K.J., Meyerowitz E.M.

J. Mol. Evol. 33:367-378(1991).

[3] Shen D., Jiang M., Hao W., Tao L., Salazar M., Fong H.K.W.

20 Biochemistry 33:13117-13125(1994).

The following descriptions of protein family functions are not provided by the Pfam or Prosite databases.

25

740. BAH

BAH domain. Number of members: 65

[1] Medline: 97074677. Molecular cloning of polybromo, a nuclear protein containing

30 multiple domains including five bromodomains, a truncated HMG-box, and two repeats of a novel domain. Nicolas RH, Goodwin GH; Gene 1996;175:233-240.

[2] Medline: 99198739. The BAH (bromo-adjacent homology) domain: a link between DNA methylation, replication and transcriptional regulation. Callebaut I, Courvalin J-C, Mornon JP; FEBS letts 1999;446:189-193.

741. ELM2.

ELM2 domain. The ELM2 (Egl-27 and MTA1 homology 2) domain is a small domain of unknown function. Number of members: 10

742. Euk proin. EUKARYOTIC_PORIN The major protein of the outer mitochondrial membrane of eukaryotes is a porin that forms a voltage-dependent anion-selective channel (VDAC) that behaves as a general diffusion pore for small hydrophilic molecules [1 to 4]. The channel adopts an open conformation at low or zero membrane potential and a closed conformation at potentials above 30-40 mV.

This protein contains about 280 amino acids and its sequence is composed of between 12 to 16 beta-strands that span the mitochondrial outer membrane. Yeast contains two members of this family (genes POR1 and POR2); vertebrates have at least three members (genes VDAC1, VDAC2 and VDAC3) [5].

A conserved region located at the C-terminal part of these proteins was selected as a signature pattern.

Consensus pattern[YH]-x(2)-D-[SPCAD]-x-[STA]-x(3)-[TAG]-[KR]-[LIVMF]-[DNSTA]-[DNS]-x(4)-[GSTAN]-[LIVMA]-x-[LIVMY]

[1] Benz R. Biochim. Biophys. Acta 1197:167-196(1994).

[2] Manella C.A. Trends Biochem. Sci. 17:315-320(1992).

[3] Dihanich M. Experientia 46:146-153(1990).

[4] Forte M., Guy H.R., Mannella C.A. J. Bioenerg. Biomembr. 19:341-350(1987).

[5] Sampson M.J., Lovell R.S., Davison D.B., Craigen W.J. Genomics 36:192-196(1996).

743. Glyco hydor 19

Chitinases family 19 signatures

cross-reference(s) CHITINASE_19_1, CHITINASE_19_2

Chitinases (EC 3.2.1.14) [1] are enzymes that catalyze the hydrolysis of the beta-1,4-N-acetyl-D-glucosamine linkages in chitin polymers. From the view point of sequence

similarity chitinases belong to either family 18 or 19 in the classification of glycosyl

hydrolases [2,E1]. Chitinases of family 19 (also known as classes IA or I and IB or II)

are enzymes from plants that function in the defense against fungal and insect pathogens

by destroying their chitin-containing cell wall. Class IA/I and IB/II enzymes differ in the

5   presence (IA/I) or absence (IB/II) of a N-terminal chitin-binding domain (see the relevant

entry <PDOC00025>). The catalytic domain of these enzymes consist of about 220 to 230

amino acid residues.

Two highly conserved regions were selected as signature patterns, the first one is located in

the N-terminal section and contains one of the six cysteines which are conserved in most,

10   if not all, of these chitinases and which is probably involved in a disulfide bond.

Consensus patternC-x(4,5)-F-Y-[ST]-x(3)-[FY]-[LIVMF]-x-A-x(3)-[YF]-x(2)-F-[GSA]

Consensus pattern[LIVM]-[GSA]-F-x-[STAG](2)-[LIVMFY]-W-[FY]-W-[LIVM]

15   [1]Flach J., Pilet P.-E., Jolles P. Experientia 48:701-716(1992).

[2] Henrissat B. Biochem. J. 280:309-316(1991).

744. MBD

20   Methyl-CpG binding domain

The Methyl-CpG binding domain (MBD) binds to DNA that contains one or more

symmetrically methylated CpGs [1]. DNA methylation in animals is associated with

alterations in chromatin structure and silencing of gene expression. MBD has negligible non-

specific affinity for DNA. In vitro foot-printing with MeCP2 showed the MBD can protect a

25   12 nucleotide region surrounding a methyl CpG pair [1]. MBDs are found in several Methyl-

CpG binding proteins and also DNA demethylase [2]. Number of members: 11

[1]Medline: 94232813. Dissection of the methyl-CpG binding domain from the chromosomal

protein MeCP2. Nan X, Meehan RR, Bird A; Nucleic Acids Res 1993;21:4886-4892.

30   [2]Medline: 99158138. A mammalian protein with specific demethylase activity for mCpG

DNA. Bhattacharya SK, Ramchandani S, Cervoni N, Szyf M; Nature 1999;397:579-583.

745. Peptidase C1

582

Eukaryotic thiol (cysteine) proteases active sites

cross-reference(s) THIOL_PROTEASE_CYS; THIOL_PROTEASE_HIS;

THIOL_PROTEASE_ASN

Eukaryotic thiol proteases (EC 3.4.22.-) [1] are a family of proteolytic enzymes which

5　contain an active site cysteine. Catalysis proceeds through a thioester intermediate and is

facilitated by a nearby histidine side chain; an asparagine completes the essential catalytic

triad. The proteases which are currently known to belong to this family are listed below

(references are only provided for recently determined sequences).

- Vertebrate lysosomal cathepsins B (EC 3.4.22.1), H (EC 3.4.22.16), L (EC 3.4.22.15),

10　and S (EC 3.4.22.27) [2].

- Vertebrate lysosomal dipeptidyl peptidase I (EC 3.4.14.1) (also known as cathepsin C)

[2].

- Vertebrate calpains (EC 3.4.22.17). Calpains are intracellular calcium- activated thiol

protease that contain both a N-terminal catalytic domain and a C-terminal calcium-binding

15　domain.

- Mammalian cathepsin K, which seems involved in osteoclastic bone resorption [3].

- Human cathepsin O [4].

- Bleomycin hydrolase. An enzyme that catalyzes the inactivation of the antitumor drug

BLM (a glycopeptide).

20　- Plant enzymes: barley aleurain (EC 3.4.22.16), EP-B1/B4; kidney bean EP-C1, rice bean

SH-EP; kiwi fruit actinidin (EC 3.4.22.14); papaya latex papain (EC 3.4.22.2),

chymopapain (EC 3.4.22.6), caricain (EC 3.4.22.30), and proteinase IV (EC 3.4.22.25);

pea turgor-responsive protein 15A; pineapple stem bromelain (EC 3.4.22.32); rape COT44;

rice oryzain alpha, beta, and gamma; tomato low-temperature induced, Arabidopsis

25　thaliana A494, RD19A and RD21A.

- House-dust mites allergens DerP1 and EurM1.

- Cathepsin B-like proteinases from the worms Caenorhabditis elegans (genes gcp-1, cpr-

3, cpr-4, cpr-5 and cpr-6), Schistosoma mansoni (antigen SM31) and Japonica (antigen

SJ31), Haemonchus contortus (genes AC-1 and AC-2), and Ostertagia ostertagi (CP-1 and

30　CP-3).

- Slime mold cysteine proteinases CP1 and CP2.

- Cruzipain from Trypanosoma cruzi and brucei.

- Throphozoite cysteine proteinase (TCP) from various Plasmodium species.

- Proteases from Leishmania mexicana, Theileria annulata and Theileria parva.

- Baculoviruses cathepsin-like enzyme (v-cath).

- Drosophila small optic lobes protein (gene sol), a neuronal protein that contains a calpain-like domain.

- Yeast thiol protease BLH1/YCP1/LAP3.

5   - Caenorhabditis elegans hypothetical protein C06G4.2, a calpain-like protein.

Two bacterial peptidases are also part of this family:

- Aminopeptidase C from Lactococcus lactis (gene pepC) [5].

10   - Thiol protease tpr from Porphyromonas gingivalis.

Three other proteins are structurally related to this family, but may have lost their proteolytic activity.

15   - Soybean oil body protein P34. This protein has its active site cysteine replaced by a glycine.

- Rat testin, a sertoli cell secretory protein highly similar to cathepsin L but with the active site cysteine is replaced by a serine. Rat testin should not be confused with mouse testin which is a LIM-domain protein (see <PDOC00382>).

20   - Plasmodium falciparum serine-repeat protein (SERA), the major blood stage antigen. This protein of 111 Kd possesses a C-terminal thiol-protease-like domain [6], but the active site cysteine is replaced by a serine.

The sequences around the three active site residues are well conserved and can be used as signature patterns.

25

Consensus patternQ-x(3)-[GE]-x-C-[YW]-x(2)-[STAGC]-[STAGCV] [C is the active site residue]

Note the residue in position 4 of the pattern is almost always cysteine; the only exceptions are calpains (Leu), bleomycin hydrolase (Ser) and yeast YCP1 (Ser). Note the residue in position

30   5 of the pattern is always Gly except in papaya protease IV where it is Glu.

Consensus pattern[LIVMGSTAN]-x-H-[GSACE]-[LIVM]-x-[LIVMAT](2)-G-x-[GSADNH] [H is the active site residue]

Consensus pattern[FYCH]-[WI]-[LIVT]-x-[KRQAG]-N-[ST]-W-x(3)-[FYW]-G-x(2)-G-[LFYW]-[LIVMFYG]-x-[LIVMF] [N is the active site residue]

584

Note these proteins belong to family C1 (papain-type) and C2 (calpains) in the classification of peptidases [7,E1].

[1]Dufour E. Biochimie 70:1335-1342(1988).

5   [2]Kirschke H., Barrett A.J., Rawlings N.D. Protein Prof. 2:1587-1643(1995).

[3]Shi G.-P., Chapman H.A., Bhairi S.M., Deleeuw C., Reddy V.Y., Weiss S.J. FEBS Lett. 357:129-134(1995).

[4]Velasco G., Ferrando A.A., Puente X.S., Sanchez L.M., Lopez-Otin C. J. Biol. Chem. 269:27136-27142(1994).

10   [5]Chapot-Chartier M.P., Nardi M., Chopin M.C., Chopin A., Gripon J.C. Appl. Environ. Microbiol. 59:330-333(1993).

[6]Higgins D.G., McConnell D.J., Sharp P.M. Nature 340:604-604(1989).

[7]Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:461-486(1994).

15

746. Peptidase M22

Glycoprotease family signature cross-reference(s) GLYCOPROTEASE

Glycoprotease (GCP) (EC 3.4.24.57) [1], or o-syaloglycoprotein endopeptidase, is a metalloprotease secreted by Pasteurella haemolytica which specifically

20   cleaves O-sialoglycoproteins such as glycophorin A. The sequence of GCP is highly similar to the following uncharacterized proteins:

- Escherichia coli hypothetical protein ygjD (ORF-X).

- Bacillus subtilis hypothetical protein ydiE.

25   - Mycobacterium leprae hypothetical protein U229E.

- Mycobacterium tuberculosis hypothetical protein MtCY78.10.

- Synechocystis strain PCC 6803 hypothetical protein slr0807.

- Methanococcus jannaschii hypothetical protein MJ1130.

- Haloarcula marismortui hypothetical protein in HSH 3'region.

30   - Yeast hypothetical protein YKR038c.

- Yeast hypothetical protein QRI7.

One of the conserved regions contains two conserved histidines. It is possible that this region is involved in coordinating a metal ion such as zinc.

Consensus pattern[KR]-[GSAT]-x(4)-[FYWLH]-[DQNGK]-x-P-x-[LIVMFY]-x(3)-H-x(2)-[AG]-H-[LIVM]

Note these proteins belong to family M22 in the classification of

5   peptidases [2,E1].

[1]Abdullah K.M., Lo R.Y.C., Mellors A. J. Bacteriol. 173:5597-5603(1991).

[2]Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).

10

747. SAM. SAM domain (Sterile alpha motif)

It has been suggested that SAM is an evolutionarily conserved protein binding domain that is involved in the regulation of numerous developmental processes in diverse eukaryotes. The SAM domain can potentially function as a protein interaction module through its ability to

15   homo- and heterooligomerise with other SAM domains. Number of members: 81

[1]Medline: 96100659 SAM: A novel motif in yeast sterile alpha and Drosophila polyhomeotic proteins Ponting CP; Prot Sci 1995;4:1928-1930.

[2]Medline: 97160498 SAM as a protein interaction domain involved in developmental

20   regulation. Shultz J, Ponting CP, Hofmann K, Bork P; Prot Sci 1997;6:249-253.

[3]Medline: 99101382 The crystal structure of an Eph receptor SAM domain reveals a mechanism for modular dimerization. Reference Author: Stapleton D, Balan I, Pawson -
T, Sicheri F; Nat Struct Biol 1999;6:44-49.

25

748. Tyrosinase signatures cross-reference(s) TYROSINASE_1; TYROSINASE_2

Tyrosinase (EC 1.14.18.1) [1] is a copper monooxygenases that catalyzes the hydroxylation of monophenols and the oxidation of o-diphenols to o-quinols.

This enzyme, found in prokaryotes as well as in eukaryotes, is involved in the

30   formation of pigments such as melanins and other polyphenolic compounds.

Tyrosinase binds two copper ions (CuA and CuB). Each of the two copper ion has been shown [2] to be bound by three conserved histidines residues. The regions around these copper-binding ligands are well conserved and also shared by some

hemocyanins, which are copper-containing oxygen carriers from the hemolymph of many molluscs and arthropods [3,4].

At least two proteins related to tyrosinase are known to exist in mammals:

5

- TRP-1 (TYRP1) [5], which is responsible for the conversion of 5,6-dihydro-xyindole-2-carboxylic acid (DHICA) to indole-5,6-quinone-2-carboxylic acid.
- TRP-2 (TYRP2) [6], which is the melanogenic enzyme DOPAchrome tautomerase (EC 5.3.3.12) that catalyzes the conversion of DOPAchrome to DHICA. TRP-2

10 differs from tyrosinases and TRP-1 in that it binds two zinc ions instead of copper [7].

Other proteins that belong to this family are:

15 - Plants polyphenol oxidases (PPO) (EC 1.10.3.1) which catalyze the oxidation of mono- and o-diphenols to o-diquinones [8].
- Caenorhabditis elegans hypothetical protein C02C2.1.

Two signature patterns for tyrosinase and related proteins have been derived

20 The first one contains two of the histidines that bind CuA, and is located in the N-terminal section of tyrosinase. The second pattern contains a histidine that binds CuB, that pattern is located in the central section of the enzyme.

Consensus pattern H-x(4,5)-F-[LIVMFTP]-x-[FW]-H-R-x(2)-[LM]-x(3)-E

25 [The two H's are copper ligands]
Consensus patternD-P-x-F-[LIVMFYW]-x(2)-H-x(3)-D [H is a copper ligand]

[1]Lerch K. Prog. Clin. Biol. Res. 256:85-98(1988).

30 [2]Jackman M.P., Hajnal A., Lerch K. Biochem. J. 274:707-713(1991).

[3]Linzen B. Naturwissenschaften 76:206-211(1989).

[4]Lang W.H., van Holde K.E. Proc. Natl. Acad. Sci. U.S.A. 88:244-248(1991).

[5]Kobayashi T., Urabe K., Winder A., Jimenez-Cervantes C., Imokawa G., Brewington T., Solano F., Garcia-Borron J.C., Hearing V.J. EMBO J. 13:5818-5825(1994).

[6]Jackson I.J., Chambers D.M., Tsukamoto K., Copeland N.G., Gilbert D.J., Jenkins N.A., Hearing V. EMBO J. 11:527-535(1992).

[7]Solano F., Martinez-Liarte J.H., Jimenez-Cervantes C., Garcia-Borron J.C., Lozano J.A. Biochem. Biophys. Res. Commun. 204:1243-1250(1994).

5   [8]Cary J.W., Lax A.R., Flurkey W.H. Plant Mol. Biol. 20:245-253(1992).

749. (Mur Ligase) Folylpolyglutamate synthase signatures

Folylpolyglutamate synthase (EC 6.3.2.17) (FPGS) [1] is the enzyme of folate metabolism

10   that catalyzes ATP-dependent addition of glutamate moieties to tetrahydrofolate.

Its sequence is moderately conserved between prokaryotes (gene folC) and eukaryotes. We developed two signature patterns based on the conserved regions which are rich in glycine residues and could play a role in the catalytical

15   activity and/or in substrate binding.

Description of pattern(s) and/or profile(s)

Consensus pattern[LIVMFY]-x-[LIVM]-[STAG]-G-T-[NK]-G-K-x-[ST]-x(7)- [LIVM](2)-x(3)-[GSK]

20   Consensus pattern[LIVMFY](2)-E-x-G-[LIVM]-[GA]-G-x(2)-D-x-[GST]-x-[LIVM](2)

[1]Shane B., Garrow T., Brenner A., Chen L., Choi Y.J., Hsu J.C., Stover P. Adv. Exp. Med. Biol. 338:629-634(1993).

25

750. (Peptidase M3) Neutral zinc metallopeptidases, zinc-binding region signature

The majority of zinc-dependent metallopeptidases (with the notable exception of the carboxypeptidases) share a common pattern of primary structure [1,2,3] in the part of their sequence involved in the binding of zinc, and can be grouped together as a

30   superfamily,known as the metzincins, on the basis of this sequence similarity. They can be classified into a number of distinct families [4,E1] which are listed below along with the proteases which are currently known to belong to these families.

Family M1

- Bacterial aminopeptidase N (EC 3.4.11.2) (gene pepN).

- Mammalian aminopeptidase N (EC 3.4.11.2).

- Mammalian glutamyl aminopeptidase (EC 3.4.11.7) (aminopeptidase A). It may play a role in regulating growth and differentiation of early B-lineage cells.

5　　- Yeast aminopeptidase yscII (gene APE2).

- Yeast alanine/arginine aminopeptidase (gene AAP1).

- Yeast hypothetical protein YIL137c.

- Leukotriene A-4 hydrolase (EC 3.3.2.6). This enzyme is responsible for the hydrolysis of an epoxide moiety of LTA-4 to form LTB-4; it has been shown that it binds zinc and is

10　　capable of peptidase activity.

Family M2

- Angiotensin-converting enzyme (EC 3.4.15.1) (dipeptidyl carboxypeptidase I) (ACE) the enzyme responsible for hydrolyzing angiotensin I to angiotensin II. There are two forms

15　　of ACE: a testis-specific isozyme and a somatic isozyme which has two active centers.

Family M3

- Thimet oligopeptidase (EC 3.4.24.15), a mammalian enzyme involved in the cytoplasmic degradation of small peptides.

20　　- Neurolysin (EC 3.4.24.16) (also known as mitochondrial oligopeptidase M or microsomal endopeptidase).

- Mitochondrial intermediate peptidase precursor (EC 3.4.24.59) (MIP). It is involved the second stage of processing of some proteins imported in the mitochondrion.

- Yeast saccharolysin (EC 3.4.24.37) (proteinase yscD).

25　　- Escherichia coli and related bacteria dipeptidyl carboxypeptidase (EC 3.4.15.5) (gene dcp).

- Escherichia coli and related bacteria oligopeptidase A (EC 3.4.24.70) (gene opdA or prlC).

- Yeast hypothetical protein YKL134c.

30　　Family M4

- Thermostable thermolysins (EC 3.4.24.27), and related thermolabile neutral proteases (bacillolysins) (EC 3.4.24.28) from various species of Bacillus.

- Pseudolysin (EC 3.4.24.26) from Pseudomonas aeruginosa (gene lasB).

- Extracellular elastase from Staphylococcus epidermidis.

- Extracellular protease prt1 from Erwinia carotovora.

- Extracellular minor protease smp from Serratia marcescens.

- Vibriolysin (EC 3.4.24.25) from various species of Vibrio.

- Protease prtA from Listeria monocytogenes.

5   - Extracellular proteinase proA from Legionella pneumophila.

Family M5

- Mycolysin (EC 3.4.24.31) from Streptomyces cacaoi.

10   Family M6

- Immune inhibitor A from Bacillus thuringiensis (gene ina). Ina degrades two classes of insect antibacterial proteins, attacins and cecropins.

Family M7

15   - Streptomyces extracellular small neutral proteases

Family M8

- Leishmanolysin (EC 3.4.24.36) (surface glycoprotein gp63), a cell surface protease from various species of Leishmania.

20

Family M9

- Microbial collagenase (EC 3.4.24.3) from Clostridium perfringens and Vibrio alginolyticus.

25   Family M10A

- Serralysin (EC 3.4.24.40), an extracellular metalloprotease from Serratia.

- Alkaline metalloproteinase from Pseudomonas aeruginosa (gene aprA).

- Secreted proteases A, B, C and G from Erwinia chrysanthemi.

- Yeast hypothetical protein YIL108w.

30

Family M10B

- Mammalian extracellular matrix metalloproteinases (known as matrixins) [5]: MMP-1 (EC 3.4.24.7) (interstitial collagenase), MMP-2 (EC 3.4.24.24) (72 Kd gelatinase), MMP-9 (EC 3.4.24.35) (92 Kd gelatinase), MMP-7 (EC 3.4.24.23) (matrylisin), MMP-8 (EC 3.4.24.34)

(neutrophil collagenase), MMP-3 (EC 3.4.24.17) (stromelysin-1), MMP-10 (EC 3.4.24.22) (stromelysin-2), and MMP-11 (stromelysin-3), MMP-12 (EC 3.4.24.65) (macrophage metalloelastase).

- Sea urchin hatching enzyme (envelysin) (EC 3.4.24.12). A protease that allows the embryo to digest the protective envelope derived from the egg extracellular matrix.

- Soybean metalloendoproteinase 1.

Family M11

- Chlamydomonas reinhardtii gamete lytic enzyme (GLE).

Family M12A

- Astacin (EC 3.4.24.21), a crayfish endoprotease.

- Meprin A (EC 3.4.24.18), a mammalian kidney and intestinal brush border metalloendopeptidase.

- Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation and which expresses metalloendopeptidase activity. The Drosophila homolog of BMP-1 is the dorsal-ventral patterning protein tolloid.

- Blastula protease 10 (BP10) from Paracentrotus lividus and the related protein SpAN from Strongylocentrotus purpuratus.

- Caenorhabditis elegans protein toh-2.

- Caenorhabditis elegans hypothetical protein F42A10.8.

- Choriolysins L and H (EC 3.4.24.67) (also known as embryonic hatching proteins LCE and HCE) from the fish Oryzias lapides. These proteases participates in the breakdown of the egg envelope, which is derived from the egg extracellular matrix, at the time of hatching.

Family M12B

- Snake venom metalloproteinases [6]. This subfamily mostly groups proteases that act in hemorrhage. Examples are: adamalysin II (EC 3.4.24.46), atrolysin C/D (EC 3.4.24.42), atrolysin E (EC 3.4.24.44), fibrolase (EC 3.4.24.72), trimerelysin I (EC 3.4.25.52) and II (EC 3.4.25.53).

- Mouse cell surface antigen MS2.

Family M13

- Mammalian neprilysin (EC 3.4.24.11) (neutral endopeptidase) (NEP).

- Endothelin-converting enzyme 1 (EC 3.4.24.71) (ECE-1), which process the precursor of endothelin to release the active peptide.

- Kell blood group glycoprotein, a major antigenic protein of erythrocytes. The Kell protein is very probably a zinc endopeptidase.

- Peptidase O from Lactococcus lactis (gene pepO).

Family M27

- Clostridial neurotoxins, including tetanus toxin (TeTx) and the various botulinum toxins (BoNT). These toxins are zinc proteases that block neurotransmitter release by proteolytic cleavage of synaptic proteins such as synaptobrevins, syntaxin and SNAP-25 [7,8].

Family M30

- Staphylococcus hyicus neutral metalloprotease.

Family M32

- Thermostable carboxypeptidase 1 (EC 3.4.17.19) (carboxypeptidase Taq), an enzyme from Thermus aquaticus which is most active at high temperature.

Family M34

- Lethal factor (LF) from Bacillus anthracis, one of the three proteins composing the anthrax toxin.

Family M35

- Deuterolysin (EC 3.4.24.39) from Penicillium citrinum and related proteases from various species of Aspergillus.

Family M36

- Extracellular elastinolytic metalloproteinases from Aspergillus.

From the tertiary structure of thermolysin, the position of the residues acting as zinc ligands and those involved in the catalytic activity are known. Two of the zinc ligands are histidines which are very close together in the sequence; C-terminal to the first histidine is

a glutamic acid residue which acts as a nucleophile and promotes the attack of a water molecule on the carbonyl carbon of the substrate. A signature pattern which includes the two histidine and the glutamic acid residues is sufficient to detect this superfamily of proteins.

5

Description of pattern(s) and/or profile(s)

Consensus pattern[GSTALIVN]-x(2)-H-E-[LIVMFYW]-{DEHRKP}-H-x-[LIVMFYWGSPQ] [The

two H's are zinc ligands] [E is the active site residue]

10   Sequences known to belong to this class detected by the patternALL,

except for members of families M5, M7 amd M11.

Other sequence(s) detected in SWISS-PROT55; including Neurospora crassa conidiation-specific protein 13 which could be a zinc-protease.

15   [1]Jongeneel C.V., Bouvier J., Bairoch A.

FEBS Lett. 242:211-214(1989).

[2]Murphy G.J.P., Murphy G., Reynolds J.J.

FEBS Lett. 289:4-7(1991).

[3]Bode W., Grams F., Reinemer P., Gomis-Rueth F.-X., Baumann U., McKay

20   D.B., Stoecker W.

Zoology 99:237-246(1996).

[4]Rawlings N.D., Barrett A.J.

Meth. Enzymol. 248:183-228(1995).

[5]Woessner J. Jr.

25   FASEB J. 5:2145-2154(1991).

[6]Hite L.A., Fox J.W., Bjarnason J.B.

[7]Montecucco C., Schiavo G.

Trends Biochem. Sci. 18:324-327(1993).

[8]Niemann H., Blasi J., Jahn R.

30   Trends Cell Biol. 4:179-185(1994).

751. PseudoU_synt_1

tRNA pseudouridine synthase is involved in the formation of pseudouridine at the anticodon stem and loop of transfer-RNAs Pseudouridine is an isomer of uridine (5-(beta-D-ribofuranosyl) uracil, and id the most abundant modified nucleoside found in all cellular RNAs. The TruA-like proteins also exhibit a conserved sequence with a strictly conserved

5   aspartic acid, likely involved in catalysis. Number of members: 25

[1]Medline: 98254513. Transfer RNA-pseudouridine synthetase Pus1 of Saccaromyces cerevisiae contains one atom of zinc essential for its native conformation and tRNA recognition. Arluison V, Hountondji C, Robert B, Grosjean H; Biochemistry 1998;37:7268-

10 7276.

752. EPSP synthase signatures

EPSP synthase (3-phosphoshikimate 1-carboxyvinyltransferase) (EC 2.5.1.19) catalyzes the

15 sixth step in the biosynthesis from chorismate of the aromatic amino acids (the shikimate pathway) in bacteria (gene aroA), plants and fungi (where it is part of a multifunctional enzyme which catalyzes five consecutive steps in this pathway) [1]. EPSP synthase has been extensively studied as it is the target of the potent herbicide glyphosate which inhibits the enzyme.

20

The sequence of EPSP from various biological sources shows that the structure of the enzyme has been well conserved throughout evolution. Two conserved regions were selected as signature patterns. The first pattern corresponds to a region that is part of the active site and which is also important for the resistance to glyphosate [2]. The second pattern is located in

25 the C-terminal part of the protein and contains a conserved lysine which seems to be important for the activity of the enzyme.

Description of pattern(s) and/or profile(s)

30 Consensus pattern[LIVM]-x(2)-[GN]-N-[SA]-G-T-[STA]-x-R-x-[LIVMY]-x-[GSTA]
Consensus pattern[KR]-x-[KH]-E-[CST]-[DNE]-R-[LIVM]-x-[STA]-[LIVMC]-x(2)-[EN]-[LIVMF]-x-[KRA]-[LIVMF]-G

[1]Stallings W.C., Abdel-Megid S.S., Lim L.W., Shieh H.-S., Dayringer H.E., Leimgruber N.K., Stegeman R.A., Anderson K.S., Sikorski J.A., Padgette S.R., Kishore G.M. Proc. Natl. Acad. Sci. U.S.A. 88:5046-5050(1991).

[2]Padgette S.R., Re D.B., Gaser C.S., Eicholtz D.A., Frazier R.B., Hironaka C.M., Levine E.B., Shah D.M., Fraley R.T., Kishore G.M. J. Biol. Chem. 266:22364-22369(1991).

753. Glyco_hydro_18

Glycosyl hydrolases family 18. Number of members: 173

[1]Medline: 95219379. Crystal structure of a bacterial chitinase at 2.3 A resolution. Perrakis A, Tews I, Dauter Z, Oppenheim AB, Chet I, Wilson KS, Vorgias CE; Structure 1994;2:1169-1180.

754. Esterase

Putative esterase

This family contains Esterase D Swiss:P10768. However it is not clear if all members of the family have the same function. This family is possibly related to the COesterase family. Number of members: 36

755. (HMA) Heavy-metal-associated domain

A conserved domain of about 30 amino acid residues has been found [1] in a number of proteins that transport or detoxify heavy metals. This domain contains two conserved cysteines that could be involved in the binding of these metals. The domain has been termed Heavy-Metal-Associated (HMA). It has been found in:

-   A variety of cation transport ATPases (E1-E2 ATPases) (see <PDOC00139>). The human copper ATPAses ATP7A and ATP7B which are respectively involved in Menke's and Wilson's diseases. ATP7A and ATP7B both contain 6 tandem copies of the HMA domain. The copper ATPases CCC2 from budding yeast, copA from Enterococcus faecalis and synA from Synechococcus contain one copy of the HMA domain. The cadmium ATPases cadA from Bacillus firmus and from plasmid pI258 from Staphylococcus aureus also contain a single HMA domain, while a chromosomal Staphylococcus aureus cadA contains two copies. Other, less characterized ATPases

that contain the HMA domain are: fixI from Rhizobium meliloti, pacS from Synechococcus strain PCC 7942), Mycobacterium leprae ctpA and ctpB and Escherichia coli hypothetical protein yhhO. In all these ATPases the HMA domain(s) are located in the N-terminal section.

5   -   Mercuric reductase (EC 1.16.1.1) (gene merA) which is generally encoded by plasmids carried by mercury-resistant Gram-negative bacteria. Mercuric reductase is a class-1 pyridine nucleotide-disulphide oxidoreductase (see <PDOC00073>). There is generally one HMA domain (with the exception of a chromosomal merA from Bacillus strain RC607 which has two) in the N-terminal part of merA.

10   -   Mercuric transport protein periplasmic component (gene merP), also encoded by plasmids carried by mercury-resistant Gram-negative bacteria. It seems to be a mercury scavenger that specifically binds to one Hg(2+) ion and which passes it to the mercuric reductase via the merT protein. The N-terminal half of merP is a HMA domain.

15   -   Helicobacter pylori copper-binding protein copP.

    -   Yeast protein ATX1 [2], which could act in the transport and/or partitioning of copper.

The consensus pattern for HMA spans the complete domain.

20

Description of pattern(s) and/or profile(s)

Consensus pattern[LIVN]-x(2)-[LIVMFA]-x-C-x-[STAGCDNH]-C-x(3)-[LIVFG]-x(3)-[LIV]-x(9,11)-[IVA]-x-[LVFYS] [The two C's probably bind metals]

25   [1]Bull P.C., Cox D.W. Trends Genet. 10:246-252(1994).

   [2]Lin S.-J., Culotta V.L. Proc. Natl. Acad. Sci. U.S.A. 92:3784-3788(1995).

756. (Peptidase M10) Matrixins cysteine switch

PROSITE cross-reference(s): CYSTEINE_SWITCH

30   Mammalian extracellular matrix metalloproteinases (EC 3.4.24.-), also known as matrixins [1] (see <PDOC00129>), are zinc-dependent enzymes. They are secreted by cells in an inactive form (zymogen) that differs from the mature enzyme by the presence of an N-terminal propeptide. A highly conserved octapeptide is found two residues downstream of the C-terminal end of the propeptide. This region has been shown to be involved in

autoinhibition of matrixins [2,3]; a cysteine within the octapeptide chelates the active site zinc ion, thus inhibiting the enzyme. This region has been called the 'cysteine switch' or 'autoinhibitor region'.

A cysteine switch has been found in the following zinc proteases:

- MMP-1 (EC 3.4.24.7) (interstitial collagenase).

- MMP-2 (EC 3.4.24.24) (72 Kd gelatinase).

- MMP-3 (EC 3.4.24.17) (stromelysin-1).

- MMP-7 (EC 3.4.24.23) (matrilysin).

- MMP-8 (EC 3.4.24.34) (neutrophil collagenase).

- MMP-9 (EC 3.4.24.35) (92 Kd gelatinase).

- MMP-10 (EC 3.4.24.22) (stromelysin-2).

- MMP-11 (EC 3.4.24.-) (stromelysin-3).

- MMP-12 (EC 3.4.24.65) (macrophage metalloelastase).

- MMP-13 (EC 3.4.24.-) (collagenase 3).

- MMP-14 (EC 3.4.24.-) (membrane-type matrix metalliproteinase 1).

- MMP-15 (EC 3.4.24.-) (membrane-type matrix metalliproteinase 2).

- MMP-16 (EC 3.4.24.-) (membrane-type matrix metalliproteinase 3).

- Sea urchin hatching enzyme (EC 3.4.24.12) (envelysin) [4].

- Chlamydomonas reinhardtii gamete lytic enzyme (GLE) [5].

Description of pattern(s) and/or profile(s)

Consensus patternP-R-C-[GN]-x-P-[DR]-[LIVSAPKQ] [C chelates the zinc ion]

[1]Woessner J. Jr. FASEB J. 5:2145-2154(1991).

[2]Sanchez-Lopez R., Nicholson R., Gesnel M.C., Matrisian L.M., Breathnach R. J. Biol. Chem. 263:11892-11899(1988).

[3]Park A.J., Matrisian L.M., Kells A.F., Pearson R., Yuan Z., Navre M. J. Biol. Chem. 266:1584-1590(1991).

[4]Lepage T., Gache C. EMBO J. 9:3003-3012(1990).

[5]Kinoshita T., Fukuzawa H., Shimada T., Saito T., Matsuda Y. Proc. Natl. Acad. Sci. U.S.A. 89:4693-4697(1992).

Attorney No. 275● ●91P

597

757. (Peptidase S8) Serine proteases, subtilase family, active sites

PROSITE cross-reference(s): PS00136; SUBTILASE_ASP, PS00137; SUBTILASE_HIS, PS00138; SUBTILASE_SER

Subtilases [1,2] are an extensive family of serine proteases whose catalytic activity is

5 provided by a charge relay system similar to that of the trypsin family of serine proteases but which evolved by independent convergent evolution. The sequence around the residues involved in the catalytic triad (aspartic acid, serine and histidine) are completely different from that of the analogous residues in the trypsin serine proteases and can be used as signatures specific to that category of proteases.

10 The subtilase family currently includes the following proteases:

- Subtilisins (EC 3.4.21.62), these alkaline proteases from various Bacillus species have been the target of numerous studies in the past thirty years.

- Alkaline elastase YaB from Bacillus sp. (gene ale).

- Alkaline serine exoprotease A from Vibrio alginolyticus (gene proA).

15 - Aqualysin I from Thermus aquaticus (gene pstI).

- AspA from Aeromonas salmonicida.

- Bacillopeptidase F (esterase) from Bacillus subtilis (gene bpf).

- C5A peptidase from Streptococcus pyogenes (gene scpA).

- Cell envelope-located proteases PI, PII, and PIII from Lactococcus lactis.

20 - Extracellular serine protease from Serratia marcescens.

- Extracellular protease from Xanthomonas campestris.

- Intracellular serine protease (ISP) from various Bacillus.

- Minor extracellular serine protease epr from Bacillus subtilis (gene epr).

- Minor extracellular serine protease vpr from Bacillus subtilis (gene vpr).

25 - Nisin leader peptide processing protease nisP from Lactococcus lactis.

- Serotype-specific antigene 1 from Pasteurella haemolytica (gene ssa1).

- Thermitase (EC 3.4.21.66) from Thermoactinomyces vulgaris.

- Calcium-dependent protease from Anabaena variabilis (gene prcA).

- Halolysin from halophilic bacteria sp. 172p1 (gene hly).

30 - Alkaline extracellular protease (AEP) from Yarrowia lipolytica (gene xpr2).

- Alkaline proteinase from Cephalosporium acremonium (gene alp).

- Cerevisin (EC 3.4.21.48) (vacuolar protease B) from yeast (gene PRB1).

- Cuticle-degrading protease (pr1) from Metarhizium anisopliae.

- KEX-1 protease from Kluyveromyces lactis.

- Kexin (EC 3.4.21.61) from yeast (gene KEX-2).

- Oryzin (EC 3.4.21.63) (alkaline proteinase) from Aspergillus (gene alp).

- Proteinase K (EC 3.4.21.64) from Tritirachium album (gene proK).

- Proteinase R from Tritirachium album (gene proR).

5   - Proteinase T from Tritirachium album (gene proT).

- Subtilisin-like protease III from yeast (gene YSP3).

- Thermomycolin (EC 3.4.21.65) from Malbranchea sulfurea.

- Furin (EC 3.4.21.85), neuroendocrine convertases 1 to 3 (NEC-1 to -3) and PACE4

protease from mammals, other vertebrates, and invertebrates. These proteases are involved

10   in the processing of hormone precursors at sites comprised of pairs of basic amino acid

residues [3].

- Tripeptidyl-peptidase II (EC 3.4.14.10) (tripeptidyl aminopeptidase) from Human.

- Prestalk-specific proteins tagB and tagC from slime mold [4]. Both proteins consist of two

domains: a N-terminal subtilase catalytic domain and a C-terminal ABC transporter domain

15   (see <PDOC00185>).

Description of pattern(s) and/or profile(s)

Consensus pattern[STAIV]-x-[LIVMF]-[LIVM]-D-[DSTA]-G-[LIVMFC]-x(2,3)-[DNH] [D

is the active site residue]

20   Consensus patternH-G-[STM]-x-[VIC]-[STAGC]-[GS]-x-[LIVMA]-[STAGCLV]-[SAGM]

[H is the active site residue]

Consensus patternG-T-S-x-[SA]-x-P-x(2)-[STAVC]-[AG] [S is the active site residue]

Note if a protein includes at least two of the three active site signatures, the probability of it

being a serine protease from the subtilase family is 100%

25   Note these proteins belong to family S8 in the classification of

peptidases [5,E1].

[1]Siezen R.J., de Vos W.M., Leunissen J.A.M., Dijkstra B.W. Protein Eng. 4:719-
737(1991).

30   [2]Siezen R.J. (In) Proceeding subtilisin symposium, Hamburg, (1992).

[3]Barr P.J. Cell 66:1-3(1991).

[4]Shaulsky G., Kuspa A., Loomis W.F.; Genes Dev. 9:1111-1122(1995).

[5]Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

758. (SSB) Single-strand binding protein family signatures

PROSITE cross-reference(s): PS00735; SSB_1,PS00736; SSB_2

The Escherichia coli single-strand binding protein [1] (gene ssb), also known as the helix-

5   destabilizing protein, is a protein of 177 amino acids. It binds tightly, as a homotetramer, to

single-stranded DNA (ss-DNA) and plays an important role in DNA replication,

recombination and repair.

Closely related variants of SSB are encoded in the genome of a variety of large self-

10   transmissible plasmids. SSB has also been characterized in bacteria such as Proteus mirabilis

or Serratia marcescens.

Eukaryotic mitochondrial proteins that bind ss-DNA and are probably involved in

mitochondrial DNA replication are structurally and evolutionary related to prokaryotic SSB.

15   Proteins currently known to belong to this subfamily are listed below [2].

- Mammalian protein Mt-SSB (P16).

- Xenopus Mt-SSBs and Mt-SSBr.

- Drosophila MtSSB.

- Yeast protein RIM1.

20

Two signature patterns have been developed for these proteins. The first is a conserved

region in the N-terminal section of the SSB's. The second is a centrally located region which,

in Escherichia coli SSB, is known to be involved in the binding of DNA.

25   Description of pattern(s) and/or profile(s)

Consensus pattern[LIVMF]-[NST]-[KRT]-[LIVM]-x-[LIVMF](2)-G-[NHRK]-[LIVM]-
[GST]-x-[DET]

Consensus patternT-x-W-[HY]-[RNS]-[LIVM]-x-[LIVMF]-[FY]-[NGKR]

30   [1]Meyer R.R., Laine P.S. Microbiol. Rev. 54:342-380(1990).

[2]Stroumbakis N.D., Li Z., Tolias P.P. Gene 143:171-177(1994).

759. KDPG and KHG aldolases active site signatures

PROSITE cross-reference(s): PS00159; ALDOLASE_KDPG_KHG_1, PS00160; ALDOLASE_KDPG_KHG_2

4-hydroxy-2-oxoglutarate aldolase (EC 4.1.3.16) (KHG-aldolase) catalyzes the

5   interconversion of 4-hydroxy-2-oxoglutarate into pyruvate and glyoxylate. Phospho-2-dehydro-3-deoxygluconate aldolase (EC 4.1.2.14) (KDPG-aldolase) catalyzes the interconversion of 6-phospho-2-dehydro-3-deoxy-D-gluconate into pyruvate and glyceraldehyde 3-phosphate.

10   These two enzymes are structurally and functionally related [1]. They are both homotrimeric proteins of approximately 220 amino-acid residues. They are class I aldolases whose catalytic mechanism involves the formation of a Schiff-base intermediate between the substrate and the epsilon-amino group of a lysine residue. In both enzymes, an arginine is required for catalytic activity.

15
Two signature patterns were developed for these enzymes. The first one contains the active site arginine and the second, the lysine involved in the Schiff-base formation.

Description of pattern(s) and/or profile(s)

20   Consensus patternG-[LIVM]-x(3)-E-[LIV]-T-[LF]-R [R is the active site residue]
Consensus patternG-x(3)-[LIVMF]-K-[LF]-F-P-[SA]-x(3)-G [K is involved in Schiff-base formation]

[1] Vlahos C J., Dekker E.E. J. Biol. Chem. 263:11683-11691(1988).

25
760. AP endonucleases family 1 signatures. PROSITE cross-reference(s): PS00726; AP_NUCLEASE_F1_1, PS00727; AP_NUCLEASE_F1_2, PS00728; AP_NUCLEASE_F1_3

30   DNA damaging agents such as the antitumor drugs bleomycin and neocarzinostatin or those that generate oxygen radicals produce a variety of lesions in DNA. Amongst these is base-loss which forms apurinic/apyrimidinic (AP) sites or strand breaks with atypical 3'termini. DNA repair at the AP sites is initiated by specific endonuclease cleavage of the

phosphodiester backbone. Such endonucleases are also generally capable of removing blocking groups from the 3'terminus of DNA strand breaks.

AP endonucleases can be classified into two families on the basis of sequence similarity. Family 1 groups the enzymes listed below [1].

- Escherichia coli exonuclease III (EC 3.1.11.2) (gene xthA).
- Streptococcus pneumoniae and Bacillus subtilis exonuclease A (gene exoA).
- Mammalian AP endonuclease 1 (AP1) (EC 4.2.99.18).
- Drosophila recombination repair protein 1 (gene Rrp1).
- Arabidopsis thaliana apurinic endonuclease-redox protein (gene arp).

Except for Rrp1 and arp, these enzymes are proteins of about 300 amino-acid residues. Rrp1 and arp both contain additional and unrelated sequences in their N-terminal section (about 400 residues for Rrp1 and 270 for arp).

Three signature patterns were developed for this family of enzymes. The patterns are based on the most conserved regions. The first pattern contains a glutamate which has been shown [2], in the Escherichia coli enzyme to bind a divalent metal ion such as magnesium or manganese

Consensus pattern[APF]-D-[LIVMF](2)-x-[LIVM]-Q-E-x-K [E binds a divalent metal ion]
Consensus patternD-[ST]-[FY]-R-[KH]-x(7,8)-[FYW]-[ST]-[FYW](2)
Consensus patternN-x-G-x-R-[LIVM]-D-[LIVMFYH]-x-[LV]-x-S

[1] Barzilay G., Hickson I.S. BioEssays 17:713-719(1995).
[2] Mol C.D., Kuo C.-F., Thayer M.M., Cunningham R.P., Tainer J.A. Nature 374:381-386(1995).

761. (ER)Enhancer of rudimentary signature, PROSITE cross-reference(s): PS01290; ER

The Drosophila protein 'enhancer of rudimentary' (gene (e(r)) is a small protein of 104 residues whose function is not yet clear. From an evolutionary point of view, it is highly

conserved [1] and has been found to exist in probably all multicellular eukaryotic organisms. It has been proposed that this protein plays a role in the cell cycle.

A conserved region in the central part of the protein was selected as as signaure pattern.

Consensus patternY-D-I-[SA]-x-L-[FY]-x-F-[IV]-D-x(3)-D-[LIV]-S

[1] Gelsthorpe M., Pulumati M., McCallum C., Dang-Vu K., Tsubota S.I. Gene 186:189-195(1997).

762. (ETF alpha) Electron transfer flavoprotein alpha-subunit signature, PROSITE cross-reference(s): PS00696; ETF_ALPHA

The electron transfer flavoprotein (ETF) [1,2] serves as a specific electron acceptor for various mitochondrial dehydrogenases. ETF transfers electrons to the main respiratory chain via ETF-ubiquinone oxidoreductase. ETF is an heterodimer that consist of an alpha and a beta subunit and which bind one molecule of FAD per dimer. A similar system also exists in some bacteria.

The alpha subunit of ETF is a protein of about 32 Kd which is structurally related to the bacterial nitrogen fixation protein fixB which could play a role in a redox process and feed electrons to ferredoxin.

Other related proteins are:

- Escherichia coli hypothetical protein ydiR.
- Escherichia coli hypothetical protein ygcQ.

A highly conserved region which is located in the C-terminal section was selected as a signature pattern for these proteins.

Consensus pattern [LI]-Y-[LIVM]-[AT]-x-G-[IV]-[SD]-G-x-[IV]-Q-H-x(2)-G-x(6)-[IV]-x-A-[IV]-N

603

[1] Finocchiaro G., Ikeda Y., Ito M., Tanaka K. Prog. Clin. Biol. Res. 321:637-652(1990).

[2] Tsai M.H., Saier M.H. Jr. Res. Microbiol. 146:397-404(1995).

763. (lectin c) C-type lectin domain signature and profile

PROSITE cross-reference(s): PS00615; C_TYPE_LECTIN_1, PS50041;
C_TYPE_LECTIN_2

A number of different families of proteins share a conserved domain which was first characterized in some animal lectins and which seem to function as a calcium-dependent carbohydrate-recognition domain [1,2,3]. This domain, which is known as the C-type lectin domain (CTL) or as the carbohydrate-recognition domain (CRD), consists of about 110 to 130 residues. There are four cysteines which are perfectly conserved and involved in two disulfide bonds. A schematic representation of the CTL domain is shown below.

```
        +------+

        |    |

xcxxxxcxxxxxxxCxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxCxxxxWxCxxxxCx

  |   |    |************|*

  +----+      +----------------------------------------------+
```

'C': conserved cysteine involved in a disulfide bond.

'c': optional cysteine involved in a disulfide bond.

'*': position of the pattern.

The categories of proteins, in which the CTL domain has been found, are listed below.

Type-II membrane proteins where the CTL domain is located at the C-terminal extremity of the proteins:

- Asialoglycoprotein receptors (ASGPR) (also known as hepatic lectins) [4]. The ASGPR's mediate the endocytosis of plasma glycoproteins to which the terminal sialic acid residue in their carbohydrate moieties has been removed.

- Low affinity immunoglobulin epsilon Fc receptor (lymphocyte IgE receptor), which plays an essential role in the regulation of IgE production and in the differentiation of B cells.

- Kupffer cell receptor. A receptor with an affinity for galactose and fucose, that could be involved in endocytosis.

- A number of proteins expressed on the surface of natural killer T-cells: NKG2, NKR-P1, YE1/88 (Ly-49), CD69 and on B-cells: CD72, LyB-2. The CTL- domain in these proteins is

5   distantly related to other CTL-domains; it is unclear whether they are likely to bind carbohydrates.

Proteins that consist of an N-terminal collagenous domain followed by a CTL- domain [5], these proteins are sometimes called 'collectins':

10

- Pulmonary surfactant-associated protein A (SP-A). SP-A is a calcium-dependent protein that binds to surfactant phospholipids and contributes to lower the surface tension at the air-liquid interface in the alveoli of the mammalian lung.

15   - Pulmonary surfactant-associated protein D (SP-D).

- Conglutinin, a calcium-dependent lectin-like protein which binds to a yeast cell wall extract and to immune complexes through the complement component (iC3b).

- Mannan-binding proteins (MBP) (also known as mannose-binding proteins).

20   MBP's bind mannose and N-acetyl-D-glucosamine in a calcium-dependent manner.

- Bovine collectin-43 (CL-43).

Selectins (or LEC-CAM) [6,7]. Selectins are cell adhesion molecules implicated in the

25   interaction of leukocytes with platelets or vascular endothelium. Structurally, selectins consist of a long extracellular domain, followed by a transmembrane region and a short cytoplasmic domain. The extracellular domain is itself composed of a CTL-domain, followed by an EGF-like domain and a variable number of SCR/Sushi repeats. Known selectins are:

30

- Lymph node homing receptor (also known as L-selectin, leukocyte adhesion molecule-1, (LAM-1), leu-8, gp90-mel, or LECAM-1)

- Endothelial leukocyte adhesion molecule 1 (ELAM-1, E-selectin or LECAM-2). The ligand recognized by ELAM-1 is sialyl-Lewis x.

- Granule membrane protein 140 (GMP-140, P-selectin, PADGEM, CD62, or LECAM-3). The ligand recognized by GMP-140 is Lewis x.

5   Large proteoglycans that contain a CTL-domain followed by one copy of a SCR/ Sushi repeat, in their C-terminal section:

- Aggrecan (cartilage-specific proteoglycan core protein). This proteoglycan is a major component of the extracellular matrix of cartilagenous tissues where it has a role in the resistance to compression.

10   - Brevican.

- Neurocan.

- Versican (large fibroblast proteoglycan), a large chondroitin sulfate proteoglycan that may play a role in intercellular signalling.

15   In addition to the CTL and Sushi domains, these proteins also contain, in their N-terminal domain, an Ig-like V-type region, two or four link domains (see <PDOC00955>) and up to two EGF-like repeats.

Two type-I membrane proteins:

20

- Mannose receptor from macrophages. This protein mediates the endocytosis of glycoproteins by macrophages in several recognition and uptake processes. Its extracellular section consists of a fibronectin type II domain followed by eight tandem repeats of the CTL domain.

25   - 180 Kd secretory phospholipase A2 receptor (PLA2-R). A protein whose structure is highly similar to that of the mannose receptor.

- DEC-205 receptor. This protein is used by dendritic cells and thymic epithelial cells to capture and endocytose diverse carbohydrate-binding antigens and direct them to antigen-processing cellular compartiments. DEC-

30   205 extracellular section consists of a fibronectin type II domain followed by ten tandem repeats of the CTL domain.

- Silk moth hemocytin, an humoral lectin which is involved in a self-defence mechanism. It is composed of 2 FA58C domains (see <PDOC00988>), a CTL domain, 2 VWFC domains (see <PDOC00928), and a CTCK (see <PDOC00912>).

Various other proteins that uniquely consist of a CTL domain:

5
- Invertebrate soluble galactose-binding lectins. A category to which belong a humoral lectin from a flesh fly; echinoidin, a lectin from the coelomic fluid of a sea urchin; BRA-2 and BRA-3, two lectins from the coelomic fluid of a barnacle, a lectin from the tunicate Polyandrocarpa misakiensis and a newt oviduct lectin. The physiological importance of these lectins is not yet known but they may play an important role in defense mechanisms.

10
- Pancreatic stone protein (PSP) (also known as pancreatic thread protein (PTP), or reg), a protein that might act as an inhibitor of spontaneous calcium carbonate precipitation.

- Pancreatitis associated protein (PAP), a protein that might be involved in the control of bacterial proliferation.

15
- Tetranectin, a plasma protein that binds to plasminogen and to isolated kringle 4.

- Eosinophil granule major basic protein (MBP), a cytotoxic protein.

- A galactose specific lectin from a rattlesnake.

- Two subunits of a coagulation factor IX/factor X-binding protein (IX/X-bp),
20
a snake venom anticoagulant protein which binds with factors IX and X in the presence of calcium.

- Two subunits of a phospholipase A2 inhibitor from the plasma of a snake (PLI-A and PLI-B).

- A lipopolysaccharide-binding protein (LPS-BP) from the hemolymph of a
25
cockroach [8].

- Sea raven antifreeze protein (AFP) [9].

As a signature pattern for this domain, the C-terminal region with its three conserved cysteines was selected.
30

Consensus patternC-[LIVMFYATG]-x(5,12)-[WL]-x-[DNSR]-x(2)-C-x(5,6)-[FYWLIVSTA]-[LIVMSTA]-C [The three C's are involved in disulfide bonds]

Note all CTL domains have five Trp residues before the second Cys,

with the exception of tunicate lectin and cockroach LPS-BP which have Leu.

Note this documentation entry is linked to both a signature pattern
5   and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so.

[1] Drickamer K. J. Biol. Chem. 263:9557-9560(1988).

10 [2] Drickamer K. Prog. Nucleic Acid Res. Mol. Biol. 45:207-232(1993).

[3] Drickamer K. Curr. Opin. Struct. Biol. 3:393-400(1993).

[4] Spiess M. Biochemistry 29:10009-10018(1990).

[5] Weis W.I., Kahn R., Fourme R., Drickamer K., Hendrickson W.A. Science 254:1608-1615(1991).

15 [6] Siegelman M. Curr. Biol. 1:125-128(1991).

[7] Lasky L.A. Science 238:964-969(1992).

[8] Jomori T., Natori S. J. Biol. Chem. 266:13318-13323(1991).

[9] Ng N.F.L., Hew C.-L. J. Biol. Chem. 267:16069-16075(1992).

20 764. (SRCR) Speract receptor repeated domain signature
PROSITE cross-reference(s): PS00420; SPERACT_RECEPTOR,

The receptor for the sea urchin egg peptide speract is a transmembrane glycoprotein of 500 amino acid residues [1]. Structurally it consists of a large extracellular domain of 450
25 residues, followed by a transmembrane region and a small cytoplasmic domain of 12 amino acids. The extracellular domain contains four repeats of a 115 amino acids domain. There are 17 positions that are perfectly conserved in the four repeats, among them are six cysteines, six glycines, and three glutamates.

30 Such a domain is also found, once, in the C-terminal section of mammalian macrophage scavenger receptor type I [2], a membrane glycoproteins implicated in the pathologic deposition of cholesterol in arterial walls during atherogenesis.

608

The signature pattern that was derived spans part of the N-terminal section of the domain and contains 8 of the 17 conserved residues.

Consensus patternG-x(5)-G-x(2)-E-x(6)-W-G-x(2)-C-x(3)-[FYW]-x(8)-C-x(3)-G

[1] Dangott J.J., Jordan J.E., Bellet R.A., Garbers D.L. Proc. Natl. Acad. Sci. U.S.A. 86:2128-2132(1989).

[2] Freeman M., Ashkenas J., Rees D.J., Kingsley D.M., Copeland N.G., Jenkins N.A., Krieger M. Proc. Natl. Acad. Sci. U.S.A. 87:8810-8814(1990).

765. Bac_surface_Ag

Bacterial surface antigen

This entry includes the following surface antigens; D15 antigen from H.influenzae, OMA87 from P.multocida, OMP85 from N.meningitidis and N.gonorrhoeae. Number of members: 14

[1]Medline: 95255676. The sequencing of the 80-kDa D15 protective surface antigen of Haemophilus influenzae. Flack FS, Loosmore S, Chong P, Thomas WR; Gene 1995;156:97-99.

[2] Medline: 96333354. Cloning, sequencing, expression, and protective capacity of the oma87 gene encoding the Pasteurella multocida 87-kilodalton outer membrane antigen. Ruffolo CG, Adler B; Infect Immun 1996;64:3161-3167.

766. BRCA1 C Terminus (BRCT) domain

The BRCT domain is found predominantly in proteins involved in cell cycle checkpoint functions responsive to DNA damage. It has been suggested that the Retinoblastoma protein contains a divergent BRCT domain, this has not been included in this family. The BRCT domain of XRCC1 forms a homodimer in the crystal structure Medline:99016060. This suggests that pairs of BRCT domains associate as homo- or heterodimers. Number of members: 131

[1] Medline: 96259550. BRCA1 protein products ...Functional motifs... Koonin EV, Altschul SF, Bork P; Nature Genet 1996;13:266-268.

[2] Medline: 97153217. From BRCA1 to RAP1: A widespread BRCT module closely associated with DNA repair Callebaut I, Mornon JP; Febs lett 1997;400:25-30.

[3] Medline: 97186552. A superfamily of conserved domains in DNA damage responsive cell cycle checkpoint proteins Bork P, Hofmann K, Bucher P, Neuwald AF, Altschul SF, Koonin EV; Faseb J 1997;11:68-76.

[4] Medline: 97402527. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ; Nucleic Acids Res 1997;25:3389-3402.

[5] Medline: 99016060. Structure of an XRCC1 BRCT domain: a new protein-protein interaction module. Zhang X, Morera S, Bates PA, Whitehead PC, Coffer AI, Hainbucher K, Nash RA, Sternberg MJ, Lindahl T, Freemont PS;

767. Kappa casein

Kappa-casein is a mammalian milk protein involved in a number of important physiological processes. In the gut, the ingested protein is split into an insoluble peptide (para kappa-casein) and a soluble hydrophilic glycopeptide (caseinomacropeptide). Caseinomacropeptide is responsible for increased efficiency of digestion, prevention of neonate hypersensitivity to ingested proteins, and inhibition of gastric pathogens. Number of members: 56

[1] Medline: 98072500. Nucleotide sequence evolution at the kappa-casein locus: evidence for positive selection within the family Bovidae. Ward TJ, Honeycutt RL, Derr JN; Genetics 1997;147:1863-1872.

768. Chitinases family 18 active site

PROSITE cross-reference(s) CHITINASE_18

Chitinases (EC 3.2.1.14) [1] are enzymes that catalyze the hydrolysis of the beta-1,4-N-acetyl-D-glucosamine linkages in chitin polymers. From the view point of sequence similarity chitinases belong to either family 18 or 19 in the classification of glycosyl hydrolases [2,E1]. Chitinases of family 18 (also known as classes III or V) groups a variety of proteins:

a) Chitinases from:

- Prokaryotes such as Alteromonas, Bacillus, Serratia, Streptomyces, etc.
- Plants such as Arabidopsis, cucumber, bean, tobacco, etc.

610

- Fungi such as Aphanocladium, Rhizopus, Saccharomyces, etc.

- Nematode (Brugia malayi).

- Insects (Manduca sexta).

- Baculoviruses (Autographa Californica Nuclear Polyhedrosis virus).

5

b) Other proteins:

- Hevamine, a rubber tree protein with chitinase and lysozyme activities.

- Kluyveromyces lactis killer toxin alpha subunit, which acts as a chitinase.

10   - Flavobacterium and Streptomyces endo-beta-N-acetylglucosaminidases (EC 3.2.1.96).

- Mammalian di-N-acetylchitobiase which is involved in the degradation of asparagine-linked glycoproteins.

- Human cartilage glycoprotein Gp-39.

- Jack bean concanavalin B (conB), a protein that has lost its catalytic activity.

15

Site directed mutagenesis experiments [3] and crystallographic data [4,5] have shown that a conserved glutamate is involved in the catalytic mechanism and probably acts as a proton donor. This glutamate is at the extremity of the best conserved region in these proteins.

20   Consensus pattern[LIVMFY]-[DN]-G-[LIVMF]-[DN]-[LIVMF]-[DN]-x-E [E is the active site residue]

[1] Flach J., Pilet P.-E., Jolles P. Experientia 48:701-716(1992).

[2] Henrissat B. Biochem. J. 280:309-316(1991).

25   [3] Watanabe T., Kohori K., Miyashita K., Fujii T., Sakai H., Uchida M., Tanaka H. J. Biol. Chem. 268:18567-18572(1993).

[4] Perrakis A., Tews I., Dauter Z., Oppenheim A.B., Chet I., Wilson K.S., Vorgias C.E. Structure 2:1169-1180(1994).

[5] van Scheltinga A.C.T., Kalk K.H., Beintema J.J., Dijkstra B.W. Structure 2:1181-

30   1189(1994).

769. gag_p17. gag gene protein p17 (matrix protein).

The matrix protein forms an icosahedral shell associated with the inner membrane of the mature immunodeficiency virus. Number of members: 1598

[1] Medline: 95055757. Three-dimensional structure of the human immunodeficiency virus type 1 matrix protein. Massiah MA, Starich MR, Paschall C, Summers MF, Christensen AM, Sundquist WI; J Mol Biol 1994;244:198-223.

5

770. GDA1/CD39 family of nucleoside phosphatases signature

PROSITE cross-reference(s); GDA1_CD39_NTPASE

A number of nucleoside diphosphate and triphosphate hydrolases as well as some yet uncharacterized proteins have been found to belong to the same family [1, 2]. This family

10 currently consist of:

- Yeast guanosine-diphosphatase (EC 3.6.1.42) (GDPase) (gene GDA1). GDA1 is a golgi integral membrane enzyme that catalyzes the hydrolysis of GDP to GMP.

- Potato apyrase (EC 3.6.1.5) (adenosine diphosphatase) (ADPase). Apyrase acts on both ATP and ADP to produce AMP.

15 - Mammalian vascular ATP-diphosphohydrolase (EC 3.6.1.5) (also known as lymphoid cell activation antigen CD39).

- Toxoplasma gondii nucleoside-triphosphatases (EC 3.6.1.15) (NTPase). NTPase hydrolyses various nucleoside triphosphates to produce the corresponding nucleoside mono- and diphosphates. This enzyme is secreted into the invaded host cell into the

20 parasitophorous vacuole, a specialized compartment where the parasite intracellulary resides.

- Pea nucleoside-triphosphatases (EC 3.6.1.15) (NTPase).

- Caenorhabditis elegans hypothetical protein C33H5.14.
25 - Caenorhabditis elegans hypothetical protein R07E4.4.
- Yeast chromosome V hypothetical protein YER005w.

The above uncharacterized proteins all seem to be membrane-bound.

30 All these proteins share a number of conserved domains. The best conserved of these domains have been selected. It is located in the central section of the proteins.

Consensus pattern[LIVM]-x-G-x(2)-E-G-x-[FY]-x-[FW]-[LIVA]-[TAG]-x-N-[HY]

[1] Handa M., Guidotti G. Biochem. Biophys. Res. Commun. 218:916-923(1996).

[2] Vasconcelos E.G., Ferreira S.T., de Carvalho T.M.U., de Souza W., Kettlun A.M., Mancilla M., Valenzuela M.A., Verjovski-Almeida S. J. Biol. Chem. 271:22139-22145(1996).

771. GTP cyclohydrolase I signatures

PROSITE cross-reference(s); GTP_CYCLOHYDROL_1_1, GTP_CYCLOHYDROL_1_2

GTP cyclohydrolase I (EC 3.5.4.16) catalyzes the biosynthesis of formic acid and dihydroneopterin triphosphate from GTP. This reaction is the first step in the biosynthesis of tetrahydrofolate in prokaryotes, of tetrahydrobiopterin in vertebrates, and of pteridine-containing pigments in insects.

GTP cyclohydrolase I is a protein of from 190 to 250 amino acid residues. The comparison of the sequence of the enzyme from bacterial and eukaryotic sources shows that the structure of this enzyme has been extremely well conserved throughout evolution [1].

Two conserved regions were selected as signature patterns. The first contains a perfectly conserved tetrapeptide which is part of the GTP-binding pocket [2], the second region also contains conserved residues involved in GTP-binding.

Consensus pattern[DEN]-[LIVM](2)-x(2)-[KRNQ]-[DEN]-[LIVM]-x(3)-[ST]-x-C-E- H-H
Consensus pattern[SA]-x-[RK]-x-Q-[LIVM]-Q-E-[RN]-[LI]-[TSN]

[1] Maier J., Witter K., Guetlich M., Ziegler I., Werner T., Ninnemann H. Biochem. Biophys. Res. Commun. 212:705-711(1995).

[2] Nar H., Huber R., Meining W., Schmid C., Weinkauf S., Bacher A. Structure 3:459-466(1995).

772. IlvC. Acetohydroxy acid isomeroreductase

Acetohydroxy acid isomeroreductase catalyses the conversion of acetohydroxy acids into dihydroxy valerates. This reaction is the second in the synthetic pathway of the essential branched side chain amino acids valine and isoleucine. Number of members: 29

[1] Medline: 97361822. The crystal structure of plant acetohydroxy acid isomeroreductase complexed with NADPH, two magnesium ions and a herbicidal transition state analog determined at 1.65 A resolution. Biou V, Dumas R, Cohen-Addad C, Douce R, Job D, Pebay-Peyroula E; EMBO J 1997;16:3405-3415.

5

773. Prokaryotic membrane lipoprotein lipid attachment site

PROSITE cross-reference(s); PROKAR_LIPOPROTEIN

In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The

10   peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):

- Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp).

- Escherichia coli lipoprotein-28 (gene nlpA).

15   - Escherichia coli lipoprotein-34 (gene nlpB).

- Escherichia coli lipoprotein nlpC.

- Escherichia coli lipoprotein nlpD.

- Escherichia coli osmotically inducible lipoprotein B (gene osmB).

- Escherichia coli osmotically inducible lipoprotein E (gene osmE).

20   - Escherichia coli peptidoglycan-associated lipoprotein (gene pal).

- Escherichia coli rare lipoproteins A and B (genes rplA and rplB).

- Escherichia coli copper homeostasis protein cutF (or nlpE).

- Escherichia coli plasmids traT proteins.

- Escherichia coli Col plasmids lysis proteins.

25   - A number of Bacillus beta-lactamases.

- Bacillus subtilis periplasmic oligopeptide-binding protein (gene oppA).

- Borrelia burgdorferi outer surface proteins A and B (genes ospA and ospB).

- Borrelia hermsii variable major protein 21 (gene vmp21) and 7 (gene vmp7).

- Chlamydia trachomatis outer membrane protein 3 (gene omp3).

30   - Fibrobacter succinogenes endoglucanase cel-3.

- Haemophilus influenzae proteins Pal and Pcp.

- Klebsiella pullulunase (gene pulA).

- Klebsiella pullulunase secretion protein pulS.

- Mycoplasma hyorhinis protein p37.

- Mycoplasma hyorhinis variant surface antigens A, B, and C (genes vlpABC).

- Neisseria outer membrane protein H.8.

- Pseudomonas aeruginosa lipopeptide (gene lppL).

- Pseudomonas solanacearum endoglucanase egl.

5 - Rhodopseudomonas viridis reaction center cytochrome subunit (gene cytC).

- Rickettsia 17 Kd antigen.

- Shigella flexneri invasion plasmid proteins mxiJ and mxiM.

- Streptococcus pneumoniae oligopeptide transport protein A (gene amiA).

- Treponema pallidium 34 Kd antigen.

10 - Treponema pallidium membrane protein A (gene tmpA).

- Vibrio harveyi chitobiase (gene chb).

- Yersinia virulence plasmid protein yscJ.

- Halocyanin from Natrobacterium pharaonis [4], a membrane associated copper- binding
15 protein. This is the first archaebacterial protein known to be modified in such a fashion).

From the precursor sequences of all these proteins, we derived a consensus pattern and a set of rules to identify this type of post-translational modification.

20 Consensus pattern{DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1) The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least one Lys or one Arg in the first seven positions of the sequence.

[1] Hayashi S., Wu H.C. J. Bioenerg. Biomembr. 22:451-471(1990).

25

[2]Klein P., Somorjai R.L., Lau P.C.K. Protein Eng. 2:15-20(1988).

[3]von Heijne G. Protein Eng. 2:531-534(1989).

[4]Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).

30

774. Aminoacyl-transfer RNA synthetases class-II signatures

PROSITE cross-reference(s); AA_TRNA_LIGASE_II_1; AA_TRNA_LIGASE_II_2

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein

biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are

5   widely diverse in terms of subunit size and of quaternary structure.

The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of

10   ATP and amino acid which is different to the Rossmann fold observed for the class I synthetases [7].

Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns from two of these regions have been

15   derived.

Consensus pattern[FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]

Consensus pattern[GSTALVF]-{DENQHRKP}-[GSTA]-[LIVMF]-[DE]-R-[LIVMF]-x-[LIVMSTAG]-[LIVMFY]

20

[1]Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).

[2]Delarue M., Moras D. BioEssays 15:675-687(1993).

[3]Schimmel P. Trends Biochem. Sci. 16:1-3(1991).

[4]Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

25   [5]Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).

[6]Cusack S. Biochimie 75:1077-1081(1993).

[7]Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-255(1990).

[8]Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).

30

775. X. Trans-activation protein X

This protein is found in hepadnaviruses where it is indispensable for replication. Number of members: 91

776.   Thymidylate synthase active site

Thymidylate synthase (EC 2.1.1.45) [1,2] catalyzes the reductive methylation of dUMP to dTMP with concomitant conversion of 5,10-methylenetetrahydrofolate to dihydrofolate. Thymidylate synthase plays an essential role in DNA synthesis and is an

5   important target for certain chemotherapeutic drugs.

Thymidylate synthase is an enzyme of about 30 to 35 Kd in most species except in protozoan and plants where it exists as a bifunctional enzyme that includes a dihydrofolate reductase domain.

A cysteine residue is involved in the catalytic mechanism (it covalently binds the 5,6-

10   dihydro-dUMP intermediate). The sequence around the active site of this enzyme is conserved from phages to vertebrates.

Consensus patternR-x(2)-[LIVM]-x(3)-[FW]-[QN]-x(8,9)-[LV]-x-P-C-[HAVM]-x(3)-[QMT]-[FYW]-x-[LV] [C is the active site residue]

15

[1] Benkovic S.J. Annu. Rev. Biochem. 49:227-251(1980).

[2] Ross P., O'Gara F., Condon S. Appl. Environ. Microbiol. 56:2156-2163(1990).

777.   Glycosyl hydrolases family 31 signatures

20   It has been shown [1,2,3,E1] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:

- Lysosomal alpha-glucosidase (EC 3.2.1.20) (acid maltase) is a vertebrate glycosidase active at low pH, which hydrolyzes alpha(1->4) and alpha(1->6) linkages in glycogen, maltose, and isomaltose.

25   - Alpha-glucosidase (EC 3.2.1.20) from the yeast Candida tsukunbaensis.

- Alpha-glucosidase (EC 3.2.1.20) (gene malA) from the archebacteria Sulfolobus solfataricus.

- Intestinal sucrase-isomaltase (EC 3.2.1.48 / EC 3.2.1.10) is a vertebrate membrane-bound, multifunctional enzyme complex which hydrolyzes sucrose, maltose and isomaltose. The

30   sucrase and isomaltase domains of the enzyme are homologous (41% of amino acid identity) and have most probably evolved by duplication.

- Glucoamylase 1 (EC 3.2.1.3) (glucan 1,4-alpha-glucosidase) from various fungal species.

- Yeast hypothetical protein YBR229c.

- Fission yeast hypothetical protein SpAC30D11.01c.

617

An aspartic acid has been implicated [4] in the catalytic activity of sucrase, isomaltase, and lysosomal alpha-glucosidase. The region around this active residue is highly conserved and can be used as a signature pattern. A second region, which contains two conserved cysteines, has been used as an additional signature pattern.

Consensus pattern [GF]-[LIVMF]-W-x-D-M-[NSA]-E [D is the active site residue]

Consensus pattern G-[AV]-D-[LIVMTA]-C-G-[FY]-x(3)-[ST]-x(3)-L-C-x-R-W-x(2)-[LV]-[GSA]-[SA]-F-x-P-F-x-R-[DN]

[1] Henrissat B. Biochem. J. 280:309-316(1991).

[2] Kinsella B.T., Hogan S., Larkin A., Cantwell B.A. Eur. J. Biochem. 202:657-664(1991).

[3] Naim H.Y., Niermann T., Kleinhans U., Hollenberg C.P., Strasser A.W.M. FEBS Lett. 294:109-112(1991).

[4] Hermans M.M.P., Kroos M.A., van Beeumen J., Oostra B.A., Reuser A.J.J. J. Biol. Chem. 266:13507-13512(1991).

778.   Urease signatures

Urease (EC 3.5.1.5) is a nickel-binding enzyme that catalyzes the hydrolysis of urea to carbon dioxide and ammonia [1]. Historically, it was the first enzyme to be crystallized (in 1926). It is mainly found in plant seeds, microorganisms and invertebrates. In plants, urease is a hexamer of identical chains. In bacteria [2], it consists of either two or three different subunits (alpha, beta and gamma).

Urease binds two nickel ions per subunit; four histidine, an aspartate and a carbamated-lysine serve as ligands to these metals; an additional histidine is involved in the catalytic mechanism [3].

As signatures for this enzyme, a region was selected that contains two histidine that bind one of the nickel ions and the region of the active site histidine.

Consensus pattern T-[AY]-[GA]-[GAT]-[LIVM]-D-x-H-[LIVM]-H-x(3)-P [The two H's bind nickel]

Consensus pattern [LIVM](2)-[CT]-H-[HN]-L-x(3)-[LIVM]-x(2)-D-[LIVM]-x-F-A [H is the active site residue]

[1] Takishima K., Suga T., Mamiya G. Eur. J. Biochem. 175:151-165(1988).

[2] Mobley H.L.T., Husinger R.P. Microbiol. Rev. 53:85-108(1989).

[3] Jabri E., Carr M.B., Hausinger R.P., Karplus P.A. Science 268:998-1004(1995).

779. Tyrosine specific protein phosphatases signature and profiles

Tyrosine specific protein phosphatases (EC 3.1.3.48) (PTPase) [1 to 5] are enzymes that catalyze the removal of a phosphate group attached to a tyrosine residue. These enzymes are very important in the control of cell growth, proliferation, differentiation and transformation. Multiple forms of PTPase have been characterized and can be classified into two categories: soluble PTPases and transmembrane receptor proteins that contain PTPase domain(s). The currently known PTPases are listed below:

Soluble PTPases.

- PTPN1 (PTP-1B).

- PTPN2 (T-cell PTPase; TC-PTP).

- PTPN3 (H1) and PTPN4 (MEG), enzymes that contain an N-terminal band 4.1- like domain (see <PDOC00566>) and could act at junctions between the membrane and cytoskeleton.

- PTPN5 (STEP).

- PTPN6 (PTP-1C; HCP; SHP) and PTPN11 (PTP-2C; SH-PTP3; Syp), enzymes which contain two copies of the SH2 domain at its N-terminal extremity. The Drosophila protein corkscrew (gene csw) also belongs to this subgroup.

- PTPN7 (LC-PTP; Hematopoietic protein-tyrosine phosphatase; HePTP).

- PTPN8 (70Z-PEP).

- PTPN9 (MEG2).

- PTPN12 (PTP-G1; PTP-P19).

- Yeast PTP1.

- Yeast PTP2 which may be involved in the ubiquitin-mediated protein degradation pathway.

- Fission yeast pyp1 and pyp2 which play a role in inhibiting the onset of mitosis.

- Fission yeast pyp3 which contributes to the dephosphorylation of cdc2.

- Yeast CDC14 which may be involved in chromosome segregation.

- Yersinia virulence plasmid PTPAses (gene yopH).

- Autographa californica nuclear polyhedrosis virus 19 Kd PTPase.

Dual specificity PTPases.

- DUSP1 (PTPN10; MAP kinase phosphatase-1; MKP-1); which dephosphorylates MAP kinase on both Thr-183 and Tyr-185.

- DUSP2 (PAC-1), a nuclear enzyme that dephosphorylates MAP kinases ERK1 and ERK2

5   on both Thr and Tyr residues.

- DUSP3 (VHR).

- DUSP4 (HVH2).

- DUSP5 (HVH3).

- DUSP6 (Pyst1; MKP-3).

10   - DUSP7 (Pyst2; MKP-X).

- Yeast MSG5, a PTPase that dephosphorylates MAP kinase FUS3.

- Yeast YVH1.

- Vaccinia virus H1 PTPase; a dual specificity phosphatase.

15   Receptor PTPases.

Structurally, all known receptor PTPases, are made up of a variable length extracellular domain, followed by a transmembrane region and a C-terminal catalytic cytoplasmic domain. Some of the receptor PTPases contain fibronectin type III (FN-III) repeats, immunoglobulin-like domains, MAM domains or carbonic anhydrase-like domains

20   in their extracellular region. The cytoplasmic region generally contains two copies of the PTPAse domain. The first seems to have enzymatic activity, while the second is inactive but seems to affect substrate specificity of the first. In these domains, the catalytic cysteine is generally conserved but some other, presumably important, residues are not.

25   In the following table, the domain structure of known receptor PTPases is shown:

	Extracellular			Intracellular
Ig	FN-3	CAH	MAM	PTPase

30

	Ig	FN-3	CAH	MAM	PTPase
Leukocyte common antigen (LCA) (CD45)	0	2	0	0	2
Leukocyte antigen related (LAR)	3	8	0	0	2
Drosophila DLAR	3	9	0	0	2
Drosophila DPTP	2	2	0	0	2

620

PTP-alpha (LRP)	0	0	0	0	2
PTP-beta	0	16	0	0	1
PTP-gamma	0	1	1	0	2
PTP-delta	0	>7	0	0	2
PTP-epsilon	0	0	0	0	2
PTP-kappa	1	4	0	1	2
PTP-mu	1	4	0	1	2
PTP-zeta	0	1	1	0	2

PTPase domains consist of about 300 amino acids. There are two conserved cysteines, the second one has been shown to be absolutely required for activity. Furthermore, a number of conserved residues in its immediate vicinity have also been shown to be important.

A signature pattern was derived for PTPase domains centered on the active site cysteine.

There are three profiles for PTPases, the first one spans the complete domain and is not specific to any subtype. The second profile is specific to dual-specificity PTPases and the third one to the PTP subfamily.

Consensus pattern [LIVMF]-H-C-x(2)-G-x(3)-[STC]-[STAGP]-x-[LIVMFY] [C is the active site residue]

Notethe M-phase inducer phosphatases (cdc25-type phosphatase) are tyrosine- protein phosphatases that are not structurally related to the above PTPases.

Notethis documentation entry is linked to both a signature pattern and to profiles. As profiles are much more sensitive than the pattern, you should use them if you have access to the necessary software tools to do so.

[1] Fischer E.H., Charbonneau H., Tonks N.K. Science 253:401-406(1991).
[2] Charbonneau H., Tonks N.K. Annu. Rev. Cell Biol. 8:463-493(1992).
[3] Trowbridge I.S. J. Biol. Chem. 266:23517-23520(1991).
[4] Tonks N.K., Charbonneau H. Trends Biochem. Sci. 14:497-500(1989).
[5] Hunter T. Cell 58:1013-1016(1989).

780.   Connexins signatures

Gap junctions [1] are specialized regions of the plasma membrane which consist of closely packed pairs of transmembrane channels, the connexons, through which small

621

molecules diffuse from a cell to a neighboring cell. Each connexon is composed of an hexamer of an integral membrane protein which is often referred to as connexin. In a given species there are a number of different, yet structurally related, tissue specific, forms of connexins. The types of connexins which are currently known are listed below.

5   - Connexin 56   (Cx56).

- Connexin 50   (Cx50)   (lens fiber protein MP70).

- Connexin 46   (Cx46)   (alpha-3).

- Connexin 45   (Cx45)   (alpha-6).

- Connexin 43   (Cx43)   (alpha-1).

10   - Connexin 40   (Cx40)   (alpha-5).

- Connexin 38   (Cx38)   (alpha-2).

- Connexin 37   (Cx37)   (alpha-4).

- Connexin 33   (Cx33)   (alpha-7).

- Connexin 32   (Cx32)   (beta-1).

15   - Connexin 31.1 (Cx31.1) (beta-4).

- Connexin 31   (Cx31)   (beta-3).

- Connexin 30.3 (Cx30.3) (beta-5).

- Connexin 26   (Cx26)   (beta-2).

Structurally the connexins consist of a short cytoplasmic N-terminal domain, followed
20   by four transmembrane segments that delimit two extracellular and one cytoplasmic loops; the C-terminal domain is cytoplasmic and its length is variable (from 20 residues in Cx26 to 260 residues in Cx56). The schematic representation of this structure is shown below.

```
NH2-***      ***      *************-COOH
25    **     **  **      **

      **    **   **   **  Cytoplasmic

   ---**____**_____**____**_____

      **    **   **   **  Membrane

      **    **   **   **

30 ---**____**_____**____**_____

      **    **   **   **  Extracellular

      **  **      **  **

         **         **
```

The sequences of the two extracellular loops are well conserved. In both loops there are three conserved cysteines which are involved in disulfide bonds. A signature patterns from each of these two loop regions has been built.

5   Consensus patternC-[DN]-T-x-Q-P-G-C-x(2)-V-C-[FY]-D [The three C's are involved in disulfide bonds] Consensus patternC-x(3,4)-P-C-x(3)-[LIVM]-[DEN]-C-[FY]-[LIVM]-[SA]-[KR]-P [The three C's are involved in disulfide bonds]

[1] Goodenough D.A., Goliger J.A., Paul D.L. Annu. Rev. Biochem. 65:475-502(1996).

10

781.   Gram-positive cocci surface proteins 'anchoring' hexapeptide

Surface proteins from Gram-positive cocci contains a conserved hexapeptide located a few residues downstream of a hydrophobic C-terminal membrane anchor region which is followed by a cluster of basic amino acids [1]. This structure is represented in the following

15   schematic representation:

```
+-----------------------------------------+-+--------+-+
| Variable length extracellular domain   |H| Anchor |B|
+-----------------------------------------+-+--------+-+
```

20   'H': conserved hexapeptide.

'B': cluster of basic residues.

It has been proposed that this hexapeptide sequence is responsible for a post-translational modification necessary for the proper anchoring of the proteins which bear it, to the cell wall.

25   Proteins known to contain such hexapeptide are listed below:

- Aggregation substance from streptococcus faecalis (asa1).

- C5a peptidase from Streptococcus pyogenes (scpA).

- C protein alpha-antigen from Streptococcus agalactiae (bca).

- Cell surface antigen I/II (PAC) from Streptococcus mutans.

30   - Dextranase from Streptococcus downei (dex).

- Fibronectin-binding protein from Staphylococcus aureus (fnbA).

- Fimbrial subunits from Actinomyces naeslundii and viscosus.

- IgA binding protein from Streptococcus pyogenes (arp4).

- IgA binding protein (B antigen) from Streptococcus agalactiae (bag).

- IgG binding proteins from Streptococci and Staphylococcus aureus.

- Internalin A from Listeria monocytogenes (inlA).

- M proteins from streptococci.

- Muramidase-released protein from Streptococcus suis (mrp).

5   - Nisin leader peptide processing protease from Lactococcus lactis (nisP).

- Protein A from Staphylococcus aureus.

- Trypsin-resistant surface T protein from streptococci.

- Wall-associated protein from Streptococcus mutans (wapA).

- Wall-associated serine proteinases from Lactococcus lactis.

10

Consensus patternL-P-x-T-G-[STGAVDE]

[1] Schneewind O., Jones K.F., Fischetti V.A. J. Bacteriol. 172:3310-3317(1990).

15   782.   Gamma-glutamyltranspeptidase signature

Gamma-glutamyltranspeptidase (EC 2.3.2.2) (GGT) [1] catalyzes the transfer of the gamma-glutamyl moiety of glutathione to an acceptor that may be an amino acid, a peptide or water (forming glutamate). GGT plays a key role in the gamma-glutamyl cycle, a pathway for the synthesis and degradation of glutathione. In prokaryotes and eukaryotes, it is an

20   enzyme that consists of two polypeptide chains, a heavy and a light subunit, processed from a single chain precursor. The active site of GGT is known to be located in the light subunit.

The sequences of mammalian and bacterial GGT show a number of regions of high similarity [2]. Pseudomonas cephalosporin acylases (EC 3.5.1.-) that convert 7-beta-(4-carboxybutanamido)-cephalosporanic acid (GL-7ACA) into 7-aminocephalosporanic acid

25   (7ACA) and glutaric acid are evolutionary related to GGT and also show some GGT activity [3]. Like GGT, these GL-7ACA acylases, are also composed of two subunits.

One of the conserved regions correspond to the N-terminal extremity of the mature light chains of these enzymes. This region has been used as a signature pattern.

30   Consensus patternT-[STA]-H-x-[ST]-[LIVMA]-x(4)-G-[SN]-x-V-[STA]-x-T-x-T-[LIVM]-[NE]-x(1,2)-[FY]-G

[1] Tate S.S., Meister A. Meth. Enzymol. 113:400-419(1985).

[2] Suzuki H., Kumagai H., Echigo T., Tochikura T. J. Bacteriol. 171:5169-5172(1989).

624

[3] Ishiye M., Niwa M. Biochim. Biophys. Acta 1132:233-239(1992).

783.   Ferrochelatase signature

Ferrochelatase (EC 4.99.1.1) (protoheme ferro-lyase) [1,2] catalyzes the last step in heme biosynthesis: the chelation of a ferrous ion to proto-porphyrin IX, to form protoheme.

In eukaryotes, ferrochelatase is a mitochondrial protein bound to the inner membrane, whose active site faces the mitochondrial matrix. The mature form of eukaryotic ferrochelatase is composed of about 360 amino acids. In bacteria, ferrochelatase (gene hemH) [3] is a protein of from 310 to 380 amino acids.

The human autosomal dominant disease protoporphyria is due to the reduced activity of ferrochelatase.

The signature pattern for this enzyme is based on a conserved region which contains a histidine residue which could be involved in binding iron.

Consensus pattern[LIVMF](2)-x-[ST]-x-H-[GS]-[LIVM]-P-x(4,5)-[DENQKR]-x-G-[DP]-x(1,2)-Y

[1] Labbe-Bois R. J. Biol. Chem. 265:7278-7283(1990).

[2] Brenner D.A., Frasier F. Proc. Natl. Acad. Sci. U.S.A. 88:849-853(1991).

[3] Miyamoto K., Nakahigashi K., Nishimura K., Inokuchi H. J. Mol. Biol. 219:393-398(1991).

784.   Cellulose-binding domain, bacterial type

The microbial degradation of cellulose and xylans requires several types of enzyme such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1].

Structurally, cellulases and xylanases generally consist of a catalytic domain joined to a cellulose-binding domain (CBD) by a short linker sequence rich in proline and/or hydroxy-amino acids.

The CBD of a number of bacterial cellulases has been shown to consist of about 105 amino acid residues [2]. Enzymes known to contain such a domain are:

- Endoglucanase (gene end1) from Butyrivibrio fibrisolvens.

- Endoglucanases A (gene cenA) and B (cenB) from Cellulomonas fimi.

- Exoglucanases A (gene cbhA) and B (cbhB) from Cellulomonas fimi.

- Endoglucanase E-2 (gene celB) from Thermomonospora fusca.

- Endoglucanase A (gene celA) from Microbispora bispora.

- Endoglucanases A (gene celA), B (celB) and C (celC) from Pseudomonas fluorescens.

- Endoglucanase A (gene celA) from Streptomyces lividans.

5   - Exocellobiohydrolase (gene cex) from Cellulomonas fimi.

- Xylanases A (gene xynA) and B (xynB) from Pseudomonas fluorescens.

- Arabinofuranosidase C (EC 3.2.1.55) (xylanase C) (gene xynC) from Pseudomonas fluorescens.

- Chitinase 63 (EC 3.2.1.14) from Streptomyces plicatus.

10   - Chitinase C from Streptomyces lividans.

The CBD domain is found either at the N-terminal or at the C-terminal extremity of these enzymes. As it is shown in the following schematic representation, there are two conserved cysteines in this CBD domain - one at each extremity of the domain - which have been shown [3] to be involved in a disulfide bond. There are also four conserved tryptophan residues which could be involved in the interaction of the CBD with polysaccharides.

```
      +-----------------------------------------------+
      |                           |
xCxxxxWxxxxxNxxxWxxxxxxxxWxxxxxxxxxWNxxxxxGxxxxxxxxxxCx
                 ********
```

'C': conserved cysteine involved in a disulfide bond. '*': position of the pattern.

Consensus patternW-N-[STAGR]-[STDN]-[LIVM]-x(2)-[GST]-x-[GST]-x(2)- [LIVMFT]-
25   [GA]

[1] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).

[2] Meinke A., Gilkes N.R., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Protein Seq. Data Anal. 4:349-353(1991).

[3] Gilkes N.R., Claeyssens M., Aebersold R., Henrissat B., Meinke A., Morrison H.D., Kilburn D.G., Warren R.A.J., Miller R.C. Jr. Eur. J. Biochem. 202:367-377(1991).

785.   Amidases signature

626

It has been shown [1,2,3] that several enzymes from various prokaryotic and eukaryotic organisms which are involved in the hydrolysis of amides (amidases) are evolutionary related. These enzymes are listed below.

5 - Indoleacetamide hydrolase (EC 3.5.1.-), a bacterial plasmid-encoded enzyme that catalyzes the hydrolysis of indole-3-acetamide (IAM) into indole-3-acetate (IAA), the second step in the biosynthesis of auxins from tryptophan.

- Acetamidase from Emericella nidulans (gene amdS), an enzyme which allows acetamide to be used as a sole carbon or nitrogen source.

10 - Amidase (EC 3.5.1.4) from Rhodococcus sp. N-774 and Brevibacterium sp. R312 (gene amdA). This enzyme hydrolyzes propionamides efficiently, and also at a lower efficiency, acetamide, acrylamide and indoleacetamide.

- Amidase (EC 3.5.1.4) from Pseudomonas chlororaphis.

- 6-aminohexanoate-cyclic-dimer hydrolase (EC 3.5.2.12) (nylon oligomers degrading enzyme E1) (gene nylA), a bacterial plasmid encoded enzyme which catalyzes the first step

15 in the degradation of 6-aminohexanoic acid cyclic dimer, a by-product of nylon manufacture [4].

- Glutamyl-tRNA(Gln) amidotransferase subunit A [5].

- Mammalian fatty acid amide hydrolase (gene FAAH) [6].

- A putative amidase from yeast (gene AMD2).

20 - Mycobacterium tuberculosis putative amidases amiA2, amiB2, amiC and amiD.

All these enzymes contain in their central section a highly conserved region rich in glycine, serine, and alanine residues. This region has been used as a signature pattern.

25 Consensus pattern: G-[GA]-S-[GS]-[GS]-G-x-[GSA]-[GSAVY]-x-[LIVM]-[GSA]-x(6)-[GSAT]-x-[GA]-x-[DE]-x-[GA]-x-S-[LIVM]-R-x-P-[GSAC]

[1] Mayaux J.-F., Cerbelaud E., Soubrier F., Faucher D., Petre D. J. Bacteriol. 172:6764-6773(1990).

30 [2] Hashimoto Y., Nishiyama M., Ikehata O., Horinouchi S., Beppu T. Biochim. Biophys. Acta 1088:225-233(1991).

[3] Chang T.-H., Abelson J. Nucleic Acids Res. 18:7180-7180(1990).

[4] Tsuchiya K., Fukuyama S., Kanzaki N., Kanagawa K., Negoro S., Okada H. J. Bacteriol. 171:3187-3191(1989).

[5] Curnow A.W., Hong K.W., Yuan R., Kim S.I., Martins O., Winkler W., Henkin T.M., Soll D. Proc. Natl. Acad. Sci. U.S.A. 94:11819-11826(1997).

[6] Cravatt B.F., Giang D.K., Mayfield S.P., Boger D.L., Lerner R.A., Gilula N.B. Nature 384:83-87(1996).

5

786.   Glycosyl hydrolases family 10 active site

The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1,2]. Fungi and bacteria produces a spectrum of cellulolytic enzymes

10   (cellulases) and xylanases which, on the basis of sequence similarities, can be classified into families. One of these families is known as the cellulase family F [3] or as the glycosyl hydrolases family 10 [4,E1]. The enzymes which are currently known to belong to this family are listed below.

- Aspergillus awamori xylanase A (xynA).

15   - Bacillus sp. strain 125 xylanase (xynA).

- Bacillus stearothermophilus xylanase.

- Butyrivibrio fibrisolvens xylanases A (xynA) and B (xynB).

- Caldocellum saccharolyticum bifunctional endoglucanase/exoglucanase (celB). This protein consists of two domains; it is the N-terminal domain, which has exoglucanase

20   activity, which belongs to this family.

- Caldocellum saccharolyticum xylanase A (xynA).

- Caldocellum saccharolyticum ORF4. This hypothetical protein is encoded in the xynABC operon and is probably a xylanase.

- Cellulomonas fimi exoglucanase/xylanase (cex).

25   - Clostridium stercorarium thermostable celloxylanase.

- Clostridium thermocellum xylanases Y (xynY) and Z (xynZ).

- Cryptococcus albidus xylanase.

- Penicillium chrysogenum xylanase (gene xylP).

- Pseudomonas fluorescens xylanases A (xynA) and B (xynB).

30   - Ruminococcus flavefaciens bifunctional xylanase XYLA (xynA). This protein consists of three domains: a N-terminal xylanase catalytic domain that belongs to family 11 of glycosyl hydrolases; a central domain composed of short repeats of Gln, Asn an Trp, and a C-terminal xylanase catalytic domain that belongs to family 10 of glycosyl hydrolases.

- Streptomyces lividans xylanase A (xlnA).

628

- Thermoanaerobacter saccharolyticum endoxylanase A (xynA).

- Thermoascus aurantiacus xylanase.

- Thermophilic bacterium Rt8.B4 xylanase (xynA).

5   One of the conserved regions in these enzymes is centered on a conserved glutamic acid residue which has been shown [5], in the exoglucanase from Cellulomonas fimi, to be directly involved in glycosidic bond cleavage by acting as a nucleophile. This region has been used as a signature pattern.

10   Consensus pattern[GTA]-x(2)-[LIVN]-x-[IVMF]-[ST]-E-[LIY]-[DN]-[LIVMF] [E is the active site residue]

[1] Beguin P. Annu. Rev. Microbiol. 44:219-248(1990).

[2] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev.
15   55:303-315(1991).

[3] Henrissat B., Claeyssens M., Tomme P., Lemesle L., Mornon J.-P. Gene 81:83-95(1989).

[4] Henrissat B. Biochem. J. 280:309-316(1991).

[5] Tull D., Withers S.G., Gilkes N.R., Kilburn D.G., Warren R.A.J., Aebersold R. J. Biol. Chem. 266:15621-15625(1991).

20

787.   Fructose-bisphosphate aldolase class-II signatures

Fructose-bisphosphate aldolase (EC 4.1.2.13) [1,2] is a glycolytic enzyme that catalyzes the reversible aldol cleavage or condensation of fructose-1,6- bisphosphate into dihydroxyacetone-phosphate and glyceraldehyde 3-phosphate. There are two classes of
25   fructose-bisphosphate aldolases with different catalytic mechanisms. Class-II aldolases [2], mainly found in prokaryotes and fungi, are homodimeric enzymes which require a divalent metal ion – generally zinc - for their activity.

This family also includes the following proteins:
30   - Escherichia coli galactitol operon protein gatY which catalyzes the transformation of tagatose 1,6-bisphosphate into glycerone phosphate and D- glyceraldehyde 3-phosphate.

- Escherichia coli N-acetyl galactosamine operon protein agaY which catalyzes the same reaction as that of gatY.

As signature patterns for this class of enzyme, two conserved regions were selected. The first pattern is located in the first half of the sequence and contains two histidine residues that have been shown [4] to be involved in binding a zinc ion. The second is located in the C-terminal section and contains clustered acidic residues and glycines.

Consensus pattern[FYVMT]-x(1,3)-[LIVMH]-[APN]-[LIVM]-x(1,2)-[LIVM]-H-x-D-H-[GACH] [The two H's are zinc ligands]

Consensus pattern[LIVM]-E-x-E-[LIVM]-G-x(2)-[GM]-[GSTA]-x-E

[1] Perham R.N. Biochem. Soc. Trans. 18:185-187(1990).

[2] Marsh J.J., Lebherz H.G. Trends Biochem. Sci. 17:110-113(1992).

[3] von der Osten C.H., Barbas C.F. III, Wong C.-H., Sinskey A.J. Mol. Microbiol. 3:1625-1637(1989).

[4] Berry A., Marshall K.E. FEBS Lett. 318:11-16(1993).

788.   Prolyl oligopeptidase family serine active site

The prolyl oligopeptidase family [1,2,3] consist of a number of evolutionary related peptidases whose catalytic activity seems to be provided by a charge relay system similar to that of the trypsin family of serine proteases, but which evolved by independent convergent evolution. The known members of this family are listed below.

- Prolyl endopeptidase (EC 3.4.21.26) (PE) (also called post-proline cleaving enzyme). PE is an enzyme that cleaves peptide bonds on the C-terminal side of prolyl residues. The sequence of PE has been obtained from a mammalian species (pig) and from bacteria (Flavobacterium meningosepticum and Aeromonas hydrophila); there is a high degree of sequence conservation between these sequences.

- Escherichia coli protease II (EC 3.4.21.83) (oligopeptidase B) (gene prtB) which cleaves peptide bonds on the C-terminal side of lysyl and argininyl residues.

- Dipeptidyl peptidase IV (EC 3.4.14.5) (DPP IV). DPP IV is an enzyme that removes N-terminal dipeptides sequentially from polypeptides having unsubstituted N-termini provided that the penultimate residue is proline.

- Yeast vacuolar dipeptidyl aminopeptidase A (DPAP A) (gene: STE13) which is responsible for the proteolytic maturation of the alpha-factor precursor.

- Yeast vacuolar dipeptidyl aminopeptidase B (DPAP B) (gene: DAP2).

630

- Acylamino-acid-releasing enzyme (EC 3.4.19.1) (acyl-peptide hydrolase). This enzyme catalyzes the hydrolysis of the amino-terminal peptide bond of an N-acetylated protein to generate a N-acetylated amino acid and a protein with a free amino-terminus.

5   A conserved serine residue has experimentally been shown (in E.coli protease II as well as in pig and bacterial PE) to be necessary for the catalytic mechanism. This serine, which is part of the catalytic triad (Ser, His, Asp), is generally located about 150 residues away from the C-terminal extremity of these enzymes (which are all proteins that contains about 700 to 800 amino acids).

10

Consensus patternD-x(3)-A-x(3)-[LIVMFYW]-x(14)-G-x-S-x-G-G-[LIVMFYW](2) [S is the active site residue]

Note these proteins belong to families S9A/S9B/S9C in the classification of peptidases
15   [4,E1].

[1] Rawlings N.D., Polgar L., Barrett A.J. Biochem. J. 279:907-911(1991).
[2] Barrett A.J., Rawlings N.D. Biol. Chem. Hoppe-Seyler 373:353-360(1992).
[3] Polgar L., Szabo E.
20   Biol. Chem. Hoppe-Seyler 373:361-366(1992).
[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

789.   Formate--tetrahydrofolate ligase signatures

Formate--tetrahydrofolate ligase (EC 6.3.4.3) (formyltetrahydrofolate synthetase)
25   (FTHFS) is one of the enzymes participating in the transfer of one-carbon units, an essential element of various biosynthetic pathways. In many of these processes the transfers of one-carbon units are mediated by the coenzyme tetrahydrofolate (THF). Various reactions generate one-carbon derivatives of THF which can be interconverted between different oxidation states by FTHFS, methylenetetrahydrofolate dehydrogenase (EC 1.5.1.5) and
30   methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9).

In eukaryotes the FTHFS activity is expressed by a multifunctional enzyme, C-1-tetrahydrofolate synthase (C1-THF synthase), which also catalyzes the dehydrogenase and cyclohydrolase activities. Two forms of C1-THF synthases are known [1], one is located in the mitochondrial matrix, while the second one is cytoplasmic. In both forms the FTHFS

domain consist of about 600 amino acid residues and is located in the C-terminal section of C1-THF synthase. In prokaryotes FTHFS activity is expressed by a monofunctional homotetrameric enzyme of about 560 amino acid residues [2].

The sequence of FTHFS is highly conserved in all forms of the enzyme. As signature patterns, two regions that are almost perfectly conserved were selected. The first one is a glycine-rich segment located in the N-terminal part of FTHFS and which could be part of an ATP-binding domain [2]. The second pattern is located in the central section of FTHFS.

Consensus patternG-[LIVM]-K-G-G-A-A-G-G-G-Y
Consensus patternV-A-T-[IV]-R-A-L-K-x-[HN]-G-G

[1] Shannon K.W., Rabinowitz J.C. J. Biol. Chem. 263:7717-7725(1988).
[2] Lovell C.R., Przybyla A., Ljungdahl L.G. Biochemistry 29:5687-5694(1990).

790.   Transthyretin signatures

Transthyretin (prealbumin) [1] is a thyroid hormone-binding protein that seems to transport thyroxine (T4) from the bloodstream to the brain. It is a protein of about 130 amino acids that assembles as a homotetramer and forms an internal channel that binds thyroxine. Transthyretin is mainly synthesized in the brain choroid plexus. In humans, variants of the protein are associated with distinct forms of amyloidosis.

The sequence of transthyretin is highly conserved in vertebrates. A number of uncharacterized proteins also belong to this family:
- Escherichia coli hypothetical protein yedX.
- Bacillus subtilis hypothetical protein yunM.
- Caenorhabditis elegans hypothetical protein R09H10.3.
- Caenorhabditis elegans hypothetical protein ZK697.8.

Two regions were selected as signature patterns. The first located in the N-terminal extremity starts with a lysine known to be involved in binding T4. The second pattern is located in the C-terminal extremity.

Consensus pattern[KH]-[IV]-L-[DN]-x(3)-G-x-P-A-x(2)-[IV]-x-[IV] [The K binds thyroxine]
Consensus patternY-[TH]-[IV]-[AP]-x(2)-L-S-[PQ]-[FYW]-[GS]-[FY]-[QS]

[1] Schreiber G., Richardson S.J. Comp. Biochem. Physiol. 116B:137-160(1997).

791.   Dihydropteroate synthase signatures

All organisms require reduced folate cofactors for the synthesis of a variety of metabolites. Most microorganisms must synthesize folate de novo because they lack the active transport system of higher vertebrate cells which allows these organisms to use dietary folates. Enzymes that are involved in the biosynthesis of folates are therefore the target of a variety of antimicrobial agents such as trimethoprim or sulfonamides.

Dihydropteroate synthase (EC 2.5.1.15) (DHPS) catalyzes the condensation of 6-hydroxymethyl-7,8-dihydropteridine pyrophosphate to para-aminobenzoic acid to form 7,8-dihydropteroate. This is the second step in the three steps pathway leading from 6-hydroxymethyl-7,8-dihydropterin to 7,8-dihydrofolate. DHPS is the target of sulfonamides which are substrates analog that compete with para-aminobenzoic acid.

Bacterial DHPS (gene sul or folP) [1] is a protein of about 275 to 315 amino acid residues which is either chromosomally encoded or found on various antibiotic resistance plasmids. In the lower eukaryote Pneumocystis carinii, DHPS is the C-terminal domain of a multifunctional folate synthesis enzyme (gene fas) [2].

Two signature patterns for DHPS were developed, the first signature is located in the N-terminal section of these enzymes, while the second signature is located in the central section.

Consensus pattern[LIVM]-x-[AG]-[LIVMF](2)-N-x-T-x-D-S-F-x-D-x-[SG]
Consensus pattern[GE]-[SA]-x-[LIVM](2)-D-[LIVM]-G-[GP]-x(2)-[STA]-x-P

[1] Slock J., Stahly D.P., Han C.-Y., Six E.W., Crawford I.P. J. Bacteriol. 172:7211-7226(1990).

[2] Volpes F., Dyer M., Scaife J.G., Darby G., Stammers D.K., Delves C.J. Gene 112:213-218(1992).

792.   Phosphatidylinositol 3- and 4-kinases signatures

Phosphatidylinositol 3-kinase (PI3-kinase) (EC 2.7.1.137) [1] is an enzyme that phosphorylates phosphoinositides on the 3-hydroxyl group of the inositol ring. The exact function of the three products of PI3-kinase - PI-3-P, PI-3,4-P(2) and PI-3,4,5-P(3) - is not

yet known, although it is proposed that they function as second messengers in cell signalling. Currently, three forms of PI3-kinase are known:

- The mammalian enzyme which is a heterodimer of a 110 Kd catalytic chain (p110) and an 85 Kd subunit (p85) which allows it to bind to activated tyrosine protein kinases. There are at

5 least two different types of p100 subunits (alpha and beta).

- Yeast TOR1/DRR1 and TOR2/DRR2 [2], PI3-kinases required for cell cycle activation. Both are proteins of about 280 Kd.

- Yeast VPS34 [3], a PI3-kinase involved in vacuolar sorting and segregation. VPS34 is a protein of about 100 Kd.

10 - Arabidopsis thaliana and soybean VPS34 homologs.

Phosphatidylinositol 4-kinase (PI4-kinase) (EC 2.7.1.67) [4] is an enzyme that acts on phosphatidylinositol (PI) in the first committed step in the production of the second messenger inositol-1,4,5,-trisphosphate. Currently the following forms of PI4-kinases are

15 known:

- Human PI4-kinase alpha.

- Yeast PIK1, a nuclear protein of 120 Kd.

- Yeast STT4, a protein of 214 Kd.

20 The PI3- and PI4-kinases share a well conserved domain at their C-terminal section; this domain seems to be distantly related to the catalytic domain of protein kinases [2]. Two signature patterns were developed from the best conserved parts of this domain.

Four additional proteins belong to this family:

25 - Mammalian FKBP-rapamycin associated protein (FRAP) [5], which acts as the target for the cell-cycle arrest and immunosuppressive effects of the FKBP12-rapamycin complex.

- Yeast protein ESR1 [6] which is required for cell growth, DNA repair and meiotic recombination.

- Yeast protein TEL1 which is involved in controlling telomere length.

30 - Yeast hypothetical protein YHR099w, a distantly related member of this family.

- Fission yeast hypothetical protein SpAC22E12.16C.

Consensus pattern[LIVMFAC]-K-x(1,3)-[DEA]-[DE]-[LIVMC]-R-Q-[DE]-x(4)-Q

Consensus pattern[GS]-x-[AV]-x(3)-[LIVM]-x(2)-[FYH]-[LIVM](2)-x-[LIVMF]-x-D-R-H-x(2)-N

[1] Hiles I.D., Otsu M., Volinia S., Fry M.J., Gout I., Dhand R., Panayotou G., Ruiz-Larrea
5   F., Thompson A., Totty N.F., Hsuan J.J., Courtneidge S.A., Parker P.J., Waterfield M.D. Cell
70:419-429(1992).

[2] Kunz J., Henriquez R., Schneider U., Deuter-Reinhard M., Movva N., Hall M.N. Cell
73:585-596(1993).

[3] Schu P.V., Takegawa K., Fry M.J., Stack J.H., Waterfield M.D., Emr S.D. Science
10   260:88-91(1993).

[4] Garcia-Bustos J.F., Marini F., Stevenson I., Frei C., Hall M.N. EMBO J. 13:2352-
2361(1994).

[5] Brown E.J., Albers M.W., Shin T.B., Ichikawa K., Keith C.T., Lane W.S., Schreiber S.L.
Nature 369:756-758(1994).

15   [6] Kato R., Ogawa H. Nucleic Acids Res. 22:3104-3112(1994).

793.   FAD-dependent glycerol-3-phosphate dehydrogenase signatures

FAD-dependent glycerol-3-phosphate dehydrogenase (EC 1.1.99.5) (GPD) catalyzes
the conversion of glycerol-3-phosphate into dihydroxyacetone phosphate. In bacteria [1] it is
20   associated with the utilization of glycerol coupled to respiration. In Escherichia coli, two
isozymes are known: one expressed under anaerobic conditions (gene glpA) and one in
aerobic conditions (gene glpD). In eukaryotes, a mitochondrial form of GPD participates in
the glycerol phosphate shuttle in conjunction with an NAD-dependent cytoplasmic GPD (EC
1.1.1.8) [2,3].

25   These enzymes are proteins of about 60 to 70 Kd which contain a probable FAD-
binding domain in their N-terminal extremity. The mammalian enzyme differs from the
bacterial or yeast proteins by having an EF-hand calcium-binding region (See
<PDOC00018>) in its C-terminal extremity.

Two signature patterns were developed. One based on the first half of the FAD-
30   binding domain and one which corresponds to a conserved region in the central part of these
enzymes.

Consensus pattern[IV]-G-G-G-x(2)-G-[STACV]-G-x-A-x-D-x(3)-R-G
Consensus patternG-G-K-x(2)-[GSTE]-Y-R-x(2)-A

[1] Austin D., Larson T.J. J. Bacteriol. 173:101-107(1991).

[2] Roennow B., Kielland-Brandt M.C. Yeast 9:1121-1130(1993).

[3] Brown L.J., McDonald M.J., Lehn D.A., Moran S.M. J. Biol. Chem. 269:14363-
14366(1994).

794.   NOL1/NOP2/sun family signature

The following proteins seems to be evolutionary related:

- Mammalian proliferating-cell nucleolar antigen p120 (gene NOL1) which may play a role in the regulation of the cell cycle and the increased nucleolar activity that is associated with the cell proliferation.

- Yeast nucleolar protein NOP2 (or YNA1) which could be involved in nucleolar function during the onset of growth, and in the maintenance of nucleolar structure.

- Yeast hypothetical protein YBL024w.

- Bacterial protein sun (also known as fmu).

- Escherichia coli hypothetical protein yebU.

- Mycobacterium tuberculosis hypothetical protein MtCY21B4.24.

- Methanococcus jannaschii hypothetical protein MJ0026.

NOL1 is a protein of 855 residues, NOP2 consists of 618 residues, YBL024w of 684, sun is a protein of about 430 to 450 residues and MJ026 has 274 residues. They share a conserved central domain which contains some highly conserved regions. One of these regions was selected as a signature pattern.

Consensus pattern[FV]-D-[KRA]-[LIVMA]-L-x-D-[AV]-P-C-[ST]-[GA]

795.   moaA / nifB / pqqE family signature

A number of proteins involved in the biosynthesis of metallo cofactors have been shown [1,2] to be evolutionary related. These proteins are:

- Bacterial and archebacterial protein moaA, which is involved in the biosynthesis of the molybdenum cofactor (molybdopterin; MPT).

- Arabidopsis thaliana cnx2, a protein involved in molybdopterin biosynthesis and which is highlys similar to moaA.

- Bacillus subtilis narA, which seems to be the moaA ortholog in that bacteria.

- Bacterial protein nifB (or fixZ) which is involved in the biosynthesis of the nitrogenase iron-molybdenum cofactor.

- Bacterial protein pqqE which is involved in the biosynthesis of the cofactor pyrrolo-quinoline-quinone (PQQ).

5   - Pyrococcus furiosus cmo, a protein involved in the synthesis of a molybdopterin-based tungsten cofactor.

- Caenorhabditis elegans hypothetical protein F49E2.1.

All these proteins share, in their N-terminal region, a conserved domain that contains three
10   cysteines. In moaA, these cysteines have been shown [1] to be important for the biological activity. They could be inolved in the binding of an iron-sulfur cluster.

Consensus pattern[LIV]-x(3)-C-[NP]-[LIVMF]-[QRS]-C-x-[FYM]-C [The three C's are putative Fe-S ligands

15

[1] Menendez C., Igloi G., Henninger H., Brandsch R. Arch. Microbiol. 164:142-151(1995).
[2] Hoff T., Schnorr K.M., Meyer C., Caboche M. J. Biol. Chem. 270:6100-6107(1995).

796.   Forkhead-associated (FHA) domain profile

20   The forkhead-associated (FHA) domain [1,E1] is a putative nuclear signalling domain found in a variety of otherwise unrelated proteins. The FHA domain comprise approximately 55 to 75 amino acids and contains three highly conserved blocks separated by divergent spacer regions. Currently it has been found in the following proteins:

- Four transcription factors that also contain a forkhead (FH) domain: mouse myocyte
25   nuclear factor 1 (MNF1), yeast transcription factor FHL1, which probably controls pre-mRNA processing, and yeast FKH1 and FKH2. In those protein the FHA domain is located N-terminal of the DNA-binding FH domain.

- Kinase-associated protein phosphatase (KAPP) from Arabidopsis thaliana, a protein which specifically interacts with the receptor-type Ser/Thr-kinase RLK5. In KAPP, the FHA
30   domain maps to a region that interacts with the receptor-type protein kinase RLK5 only if the kinase is phosphorylated on serine residues [2].

- Two protein kinases from yeast that are involved in mediating the nuclear response to DNA damage: DUN1 and SPK1/SAD1 [3]. The latter is the only known protein containing two copies of the FHA domain.

637

- Protein kinase cds1 from fission yeast contains a FHA domain and might be the ortholog of SPK1.

- Protein kinase MEK1 from yeast, which is involved in meiotic recombination.

- Human nuclear antigen Ki67 which is expressed only in proliferating cells.

5   - Yeast hypothetical protein YHR115c, which contains a RING-finger C-terminal of the FHA domain.

- Yeast hypothetical proteins L8083.1 and 9346.10, which contain an extensive coiled-coil region C-terminal of the FHA domain.

- Caenorhabditis elegans hypothetical protein ZK632.2.

10   - Caenorhabditis elegans hypothetical protein C01G6.5.

- FraH from the prokaryote Anabaena, which contains a zinc-finger motif N-terminal of the FHA domain.

- An ORF from the bacterium Streptomyces, which is on the opposite strand of the protein kinase pks1, overlapping the ORF of the kinase.

15

[1] Hofmann K.O., Bucher P. Trends Biochem. Sci. 20:347-349(1995).

[2] Stone J.M., Collinge M.A., Smith R.D., Horn M.A., Walker J.C. Science 266:793-795(1994).

[3] Navas T.A., Zhou Z., Elledge S.J. Cell 80:29-39(1995).

20

797.   Ald_Xan_dh_C

Aldehyde oxidase and xanthine dehydrogenase, C terminus

[1] Romao MJ, Archer M, Moura I, Moura JJ, LeGall J, Engh R, Schneider M, Hof P, Huber
25   R; Medline: 96072968 "Crystal structure of the xanthine oxidase-related aldehyde oxido-reductase from D. gigas." Science 1995;270:1170-1176.

Number of members: 54

30   798.   Glyco_hydro_38
Glycosyl hydrolases family 38

Glycosyl hydrolases are key enzymes of carbohydrate metabolism.

Number of members: 20

[1] Henrissat B; Medline: 98313424; Glycosidase families" Biochem Soc Trans 1998;26:153-156.

5   799.   HECT

HECT-domain (ubiquitin-transferase).

The name HECT comes from Homologous to the E6-AP Carboxyl Terminus.

10   Number of members: 43

[1] Huibregtse JM, Scheffner M, Beaudenon S, Howley PM; Medline: 95223981; A family of proteins structurally and functionally related to the E6-AP ubiquitin-protein ligase." Proc Natl Acad Sci U S A 1995;92:2563-2567.

15   800.   HRDC

HRDC domain

The HRDC (Helicase and RNase D C-terminal) domain has a putative role in nucleic acid binding. Mutations in the HRDC domain cause human disease.

20   Number of members: 19

[1] Morozov V, Mushegian AR, Koonin EV, Bork P; Medline: 98060076; A putative nucleic acid-binding domain in Bloom's and Werner's syndrome helicases" Trends Biochem
25   Sci 1997;22:417-418.

801.   Integrase

Integrase mediates integration of a DNA copy of the viral genome into the host chromosome. Integrase is composed of three domains. The amino-terminal domain is a zinc
30   binding domain. The central domain is the catalytic domain [1].The carboxyl terminal domain is a DNA binding domain [2].

Number of members: 581

[1] Dyda F, Hickman AB, Jenkins TM, Engelman A, Craigie R, Davies DR; Medline: 95099322. Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases." Science 1994;266:1981-1986.

[2] Lodi PJ, Ernst JA, Kuszewski J, Hickman AB, Engelman A, Craigie R, Clore GM,
5   Gronenborn AM; Medline: 95359147;   Solution structure of the DNA binding domain of HIV-1 integrase." Biochemistry 1995;34:9826-9833

802.   lig_chan

Ligand-gated ion channel

10   This family includes the four transmembrane regions of the ionotropic glutamate receptors and NMDA receptors.

Number of members: 128

15   [1] Tong G, Shepherd D, Jahr CE; Medline: 95184014;   Synaptic desensitization of NMDA receptors by calcineurin." Science 1995;267:1510-1512.

803.   RhoGAP

RhoGAP domain

20   GTPase activator proteins towards Rho/Rac/Cdc42-like small GTPases.

Number of members: 97

[1] Musacchio A, Cantley LC, Harrison SC; Medline: 97121392;   Crystal structure of the
25   breakpoint cluster region-homology domain from phosphoinositide 3-kinase p85 alpha subunit." Proc Natl Acad Sci U S A 1996;93:14373-14378.

[2] Barrett T, Xiao B, Dodson EJ, Dodson G, Ludbrook SB, Nurmahomed K, Gamblin SJ, Musacchio A, Smerdon SJ, Eccleston JF; Medline: 97162209;   The structure of the GTPase-activating domain from p50rhoGAP." Nature 1997;385:458-461.

30   [3] Rittinger K, Walker PA, Eccleston JF, Nurmahomed K, Owen D, Laue E, Gamblin SJ, Smerdon SJ; Medline: 97404320;   Crystal structure of a small G protein in complex with the GTPase-activating protein rhoGAP." Nature 1997;388:693-697.

[4] Boguski MS, McCormick F; Medline: 94081948;   Proteins regulating Ras and its relatives." Nature 1993;366:643-654.

804.   vwd

von Willebrand factor type D domain

5   [1] Bork P; Medline: 93327926;   The modular architecture of a new family of growth

regulators related to connective tissue growth factor." FEBS lett 1993;327:125-130.

Number of members: 92

10   805.   zf-C4_Topoisom

Topoisomerase DNA binding C4 zinc finger

[1] Tse-Dinh YC, Beran-Steed RK; Medline: 89034032;   Escherichia coli DNA

topoisomerase I is a zinc

15   metalloprotein with three repetitive zinc-binding domains." J Biol Chem 1988;263:15857-

15859.

[2] Ahumada A, Tse-Dinh YC; Medline: 99011409;   The Zn(II) binding motifs of E. coli

DNA topoisomerase I is part of a high-affinity DNA binding domain." Biochem Biophys Res

Commun 1998;251:509-514.

20

Number of members: 51

806. AIRC

AIR carboxylase

25   Members of this family catalyse the decarboxylation of 1-(5-phosphoribosyl)-5-amino-4-

imidazole-carboxylate (AIR). This family catalyse the sixth step of de novo purine

biosynthesis. Some members of this family contain two copies of this

domain. Number of members: 35

30   807. Bromodomain signature and profile

PROSITE cross-reference(s): PS00633; BROMODOMAIN_1, PS50014;

BROMODOMAIN_2

The bromodomain [1,2,3] is a conserved region of about 70 amino acids found in the

following proteins:

- Higher eukaryotes transcription initiation factor TFIID 250 Kd subunit (TBP-associated factor p250) (gene CCG1). P250 associated with the TFIID TATA-box binding protein and seems essential for progression of the G1 phase of the cell cycle.

5   - Human RING3, a protein of unknown function encoded in the MHC class II locus.

- Mammalian CREB-binding protein (CBP), which mediates cAMP-gene regulation by binding specifically to phosphorylated CREB protein.

- Drosophila female sterile homeotic protein (gene fsh), required maternally for proper expression of other homeotic genes involved in pattern formation, such as Ubx.

10   - Drosophila brahma protein (gene brm), a protein required for the activation of multiple homeotic genes.

- Mammalian homologs of brahma. In human, three brahma-like proteins are known: SNF2a(hBRM), SNF2b, and BRG1.

- Human BS69, a protein that binds to adenovirus E1A and inhibits E1A transactivation

15   - Human peregrin (or Br140).

- Yeast BDF1 [3], a transcription factor involved in the expression of a broad class of genes including snRNAs.

- Yeast GCN5, a general transcriptional activator operating in concert with certain other DNA-binding transcriptional activators, such as GCN4, HAP2/3/4 or ADA2.

20   - Yeast NPS1/STH1, involved in G(2) phase control in mitosis.

- Yeast SNF2/SWI2, which is part of a complex with the SNF5, SNF6, SWI3 and ADR6/SWI1 proteins. This SWI-complex is involved in transcriptional activation.

- Yeast SPT7, a transcriptional activator of Ty elements and possibly other genes.

- Caenorhabditis elegans protein cbp-1.

25   - Yeast hypothetical protein YGR056w.

- Yeast hypothetical protein YKR008w.

- Yeast hypothetical protein L9638.1.

Some proteins contain a region which, while similar to some extent to a classical

30   bromodomain, diverges from it by either lacking part of the domain or because of an insertion. These proteins are:

- Mammalian protein HRX (also known as All-1 or MLL), a protein involved in translocations leading to acute leukemias and which possibly acts as a transcriptional regulatory factor. HRX contains a region similar to the C- terminal half of the bromodomain.

- Caenorhabditis elegans hypothetical protein ZK783.4. The bromodomain of this protein has

5   a 23 amino-acid insertion.

- Yeast protein YTA7. This protein contains a region with significant similarity to the C-terminal half of the bromodomain. As it is a member of the AAA family (see <PDOC00572>) it is also in a functionally different context.

10   The above proteins generally contain a single bromodomain, but some of them contain two copies, this is the case of BDF1, CCG1, fsh, RING3, YKR008w and L9638.1.

The exact function of this domain is not yet known but it is thought to be involved in protein-protein interactions and it may be important for the assembly or activity of multicomponent

15   complexes involved in transcriptional activation.

The consensus pattern that has been developed spans a major part of the bromodomain; a more sensitive detection is available through the use of a profile which spans the whole domain.

20

Consensus pattern[STANVF]-x(2)-F-x(4)-[DNS]-x(5,7)-[DENQTF]-Y-[HFY]-x(2)-[LIVMFY]-x(3)-[LIVM]-x(4)-[LIVM]-x(6,8)-Y-x(12,13)-[LIVM]-x(2)-N-[SACF]-x(2)-[FY]

25   References

[1] Haynes S.R., Doolard C., Winston F., Beck S., Trowsdale J., Dawid I.B. Nucleic Acids Res. 20:2693-2603(1992).

[2] Tamkun J.W., Deuring R., Scott M.P., Kissinger M., Pattatucci A.M., Kaufman T.C., Kennison J.A. Cell 68:561-572(1992).

30   [3] Tamkun J.W. Curr. Opin. Genet. Dev. 5:473-477(1995).

808. (CH) Actinin-type actin-binding domain signatures

PROSITE cross-reference(s): PS00019; ACTININ_1, PS00020; ACTININ_2

Alpha-actinin is a F-actin cross-linking protein which is thought to anchoractin to a variety of intracellular structures [1]. The actin-binding domain of alpha-actinin seems to reside in the first 250 residues of the protein. A similar actin-binding domain has been found in the N-terminal region of many different actin-binding proteins [2,3]:

- In the beta chain of spectrin (or fodrin).

- In dystrophin, the protein defective in Duchenne muscular dystrophy (DMD) and which may play a role in anchoring the cytoskeleton to the plasma membrane.

- In the slime mold gelation factor (or ABP-120).

- In actin-binding protein ABP-280 (or filamin), a protein that link actin filaments to membrane glycoproteins.

- In fimbrin (or plastin), an actin-bundling protein. Fimbrin differs from the above proteins in that it contains two tandem copies of the actin-binding domain and that these copies are located in the C-terminal part of the protein.

Two conserved regions were selected as signature patterns for this type of main. The first of this region is located at the beginning of the domain, hile the second one is located in the central section and has been shown to be essential for the binding of actin.

Consensus pattern[EQ]-x(2)-[ATV]-[FY]-x(2)-W-x-N
Consensus pattern[LIVM]-x-[SGN]-[LIVM]-[DAGHE]-[SAG]-x-[DNEAG]-[LIVM]-x-[DEAG]-x(4)-[LIVM]-x-[LM]-[SAG]-[LIVM]-[LIVMT]-W-x- [LIVM](2)

[1] Schleicher M., Andre E., Harmann A., Noegel A.A. Dev. Genet. 9:521-530(1988).
[2] Matsudaira P. Trends Biochem. Sci. 16:87-92(1991).
[3] Dubreuil R.R. BioEssays 13:219-226(1991).

809. (COX1) Heme-copper oxidase subunit I, copper B binding region signature
PROSITE cross-reference(s): PS00077; COX1
Heme-copper respiratory oxidases [1] are oligomeric integral membrane protein complexes that catalyze the terminal step in the respiratory chain: they transfer electrons from cytochrome c or a quinol to oxygen. Some terminal oxidases generate a transmembrane proton gradient across the plasma membrane (prokaryotes) or the mitochondrial inner membrane (eukaryotes). The enzyme

complex consists of 3-4 subunits (prokaryotes) up to 13 polypeptides (mammals) of which only the catalytic subunit (equivalent to mammalian subunit 1 (CO I)) is found in all heme-copper respiratory oxidases. The presence of a bimetallic center (formed by a high-spin heme and copper B) as well as a low-spin heme,

5   both ligated to six conserved histidine residues near the outer side of four transmembrane spans within CO I is common to all family members [2-4].

In contrary to eukaryotes the respiratory chain of prokaryotes is branched to multiple terminal oxidases. The enzyme complexes vary in heme and copper

10 composition, substrate type and substrate affinity. The different respiratory oxidases allow the cells to customize their respiratory systems according a variety of environmental growth conditions [1].

Recently also a component of an anaerobic respiratory chain has been found to

15 contain the copper B binding signature of this family: nitric oxide reductase (NOR) exists in denitrifying species of Archae and Eubacteria.

Enzymes that belong to this family are:

20 - Mitochondrial-type cytochrome c oxidase (EC 1.9.3.1) which uses cytochrome c as electron donor. The electrons are transferred via copper A (Cu(A)) and heme a to the bimetallic center of CO I that is formed by a penta-coordinated heme a and copper B (Cu(B)). Subunit 1 contains 12 transmembrane regions. Cu(B) is said to be ligated to three of the

25 conserved histidine residues within the transmembrane segments 6 and 7.
- Quinol oxidase from prokaryotes that transfers electrons from a quinol to the binuclear center of polypeptide I. This category of enzymes includes Escherichia coli cytochrome O terminal oxidase complex which is a component of the aerobic respiratory chain that predominates when cells are grown at

30 high aeration.
- FixN, the catalytic subunit of a cytochrome c oxidase expressed in nitrogen-fixing bacteroids living in root nodules. The high affinity for oxygen allows oxidative phosphorylation under low oxygen concentrations. A similar enzyme has been found in other purple bacteria.

- Nitric oxide reductase (EC 1.7.99.7) from Pseudomonas stutzeri. NOR reduces nitrate to dinitrogen. It is a heterodimer of norC and the catalytic

subunit norB. The latter contains the 6 invariant histidine residues and 12 transmembrane segments [5].

5

As a signature pattern the copper-binding region was used.

Consensus pattern[YWG]-[LIVFYWTA](2)-[VGS]-H-[LNP]-x-V-x(44,47)-H-H [The three H's are copper B ligands]

10

Notecytochrome bd complexes do not belong to this family.

[1]
Garcia-Horsman J.A., Barquera B., Rumbley J., Ma J., Gennis R.B.
15   J. Bacteriol. 176:5587-5600(1994).
[2]
Castresana J., Luebben M., Saraste M., Higgins D.G.
EMBO J. 13:2516-2525(1994).
[3]
20   Capaldi R.A., Malatesta F., Darley-Usmar V.M.
Biochim. Biophys. Acta 726:135-148(1983).
[4]
Holm L., Saraste M., Wikstrom M.
EMBO J. 6:2819-2823(1987).
25   [5]
Saraste M., Castresana J.
FEBS Lett. 341:1-4(1994).

810. (dehydrog_molyb) Eukaryotic molybdopterin oxidoreductases signature
30   PROSITE cross-reference(s): PS00559; MOLYBDOPTERIN_EUK

A number of different eukaryotic oxidoreductases that require and bind a molybdopterin cofactor have been shown [1] to share a few regions of sequence similarity. These enzymes are:

- Xanthine dehydrogenase (EC 1.1.1.204), which catalyzes the oxidation of

xanthine to uric acid with the concomitant reduction of NAD. Structurally,

this enzyme of about 1300 amino acids consists of at least three distinct

5   domains: an N-terminal 2Fe-2S ferredoxin-like iron-sulfur binding domain

(see <PDOC00175>), a central FAD/NAD-binding domain and a C-terminal Mo-

pterin domain.

- Aldehyde oxidase (EC 1.2.3.1), which catalyzes the oxidation aldehydes into

acids. Aldehyde oxidase is highly similar to xanthine dehydrogenase in its

10   sequence and domain structure.

- Nitrate reductase (EC 1.6.6.1), which catalyzes the reduction of nitrate

to nitrite. Structurally, this enzyme of about 900 amino acids consists of

an N-terminal Mo-pterin domain, a central cytochrome b5-type heme-binding

domain (see <PDOC00170>) and a C-terminal FAD/NAD-binding cytochrome

15   reductase domain.

- Sulfite oxidase (EC 1.8.3.1), which catalyzes the oxidation of sulfite to

sulfate. Structurally, this enzyme of about 460 amino acids consists of an

N-terminal cytochrome b5-binding domain followed by a Mo-pterin domain.

20   There are a few conserved regions in the sequence of the molybdopterin-binding

domain of these enzymes. The pattern uses to detect these proteins is based

on one of them. It contains a cysteine residue which could be involved in

binding the molybdopterin cofactor.

25   Consensus pattern[GA]-x(3)-[KRNQHT]-x(11,14)-[LIVMFYWS]-x(8)-[LIVMF]-x-C-
x(2)-[DEN]-R-x(2)-[DE]

[1]

Wootton J.C., Nicolson R.E., Cock J.M., Walters D.E., Burke J.F., Doyle

30   W.A., Bray R.C.

Biochim. Biophys. Acta 1057:157-185(1991).

811. (DNA_ligase) ATP-dependent DNA ligase signatures
PROSITE cross-reference(s): PS00697; DNA_LIGASE_A1, PS00333; DNA_LIGASE_A2

DNA ligase (polydeoxyribonucleotide synthase) is the enzyme that joins two DNA fragments by catalyzing the formation of an internucleotide ester bond between phosphate and deoxyribose. It is active during DNA replication, DNA repair and

5   DNA recombination. There are two forms of DNA ligase: one requires ATP (EC 6.5.1.1), the other NAD (EC 6.5.1.2).

Eukaryotic, archaebacterial, virus and phage DNA ligases are ATP-dependent. During the first step of the joining reaction, the ligase interacts with ATP

10   to form a covalent enzyme-adenylate intermediate. A conserved lysine residue is the site of adenylation [1,2].

Apart from the active site region, the only conserved region common to all ATP-dependent DNA ligases is found [3] in the C-terminal section and contains

15   a conserved glutamate as well as four positions with conserved basic residues.

Signature patterns were developed for both conserved regions.

Consensus pattern[EDQH]-x-K-x-[DN]-G-x-R-[GACIVM] [K is the active site

20   residue]

Consensus patternE-G-[LIVMA]-[LIVM](2)-[KR]-x(5,8)-[YW]-[QNEK]-x(2,6)-[KRH]-x(3,5)-K-[LIVMFY]-K

Sequences known to belong to this class detected by the patternALL, except

25   for archebacterial DNA ligases.

[1]
Tomkinson A.E., Totty N.F., Ginsburg M., Lindahl T.
Proc. Natl. Acad. Sci. U.S.A. 88:400-404(1991).

30   [2]
Lindahl T., Barnes D.E.
Annu. Rev. Biochem. 61:251-281(1992).
[3]
Kletzin A.

Nucleic Acids Res. 20:5389-5396(1992).

812. (FAD_Gly3P_dh) FAD-dependent glycerol-3-phosphate dehydrogenase signatures

PROSITE cross-reference(s): PS00977; FAD_G3PDH_1, PS00978; FAD_G3PDH_2

5

FAD-dependent glycerol-3-phosphate dehydrogenase (EC 1.1.99.5) (GPD) catalyzes

the conversion of glycerol-3-phosphate into dihydroxyacetone phosphate. In

bacteria [1] it is associated with the utilization of glycerol coupled to

respiration. In Escherichia coli, two isozymes are known: one expressed under

10   anaerobic conditions (gene glpA) and one in aerobic conditions (gene glpD). In

eukaryotes, a mitochondrial form of GPD participates in the glycerol phosphate

shuttle in conjunction with an NAD-dependent cytoplasmic GPD (EC 1.1.1.8) [2,

3].

15   These enzymes are proteins of about 60 to 70 Kd which contain a probable

FAD-binding domain in their N-terminal extremity. The mammalian enzyme differs

from the bacterial or yeast proteins by having an EF-hand calcium-binding

region (See <PDOC00018>) in its C-terminal extremity.

20   Two signature patterns were developed. One based on the first half of the FAD-

binding domain and one which corresponds to a conserved region in the central

part of these enzymes.

Consensus pattern[IV]-G-G-G-x(2)-G-[STACV]-G-x-A-x-D-x(3)-R-G

25

Consensus patternG-G-K-x(2)-[GSTE]-Y-R-x(2)-A

[1]

Austin D., Larson T.J.

J. Bacteriol. 173:101-107(1991).

30   [2]

Roennow B., Kielland-Brandt M.C.

Yeast 9:1121-1130(1993).

[3]

Brown L.J., McDonald M.J., Lehn D.A., Moran S.M.

J. Biol. Chem. 269:14363-14366(1994).

813. (Fapy_DNA_glyco) Formamidopyrimidine-DNA glycosylase signature
PROSITE cross-reference(s): PS01242; FPG

5

Formamidopyrimidine-DNA glycosylase (EC 3.2.2.23) [1] (Fapy-DNA glycosylase)
(gene fpg) is a bacterial enzyme involved in DNA repair and which excise
oxidized purine bases to release 2,6-diamino-4-hydroxy-5N-methylformamido-
pyrimidine (Fapy) and 7,8-dihydro-8-oxoguanine (8-OxoG) residues. In addition

10 to its glycosylase activity, FPG can also nick DNA at apurinic/apyrimidinic
sites (AP sites). FPG is a monomeric protein of about 32 Kd which binds and
require zinc for its activity.

The binding site for zinc seems to be located in the C-terminal part of the

15 enzyme where fours conserved and essential [2] cysteines are located. A signature pattern
was developed based on this region.

Consensus patternC-x(2,4)-C-x-[GTAQ]-x-[IV]-x(7)-R-[GSTAN]-[STA]-x-[FYI]-C- x(2)-C-
Q

20 [The four C's are putative zinc ligands]

[1]
Duwat P., de Oliveira R., Ehrlich S.D., Boiteux S.
Microbiology 141:411-417(1995).

25 [2]
O'Connor T.E., Graves R.J., Demurcia G., Castaing B., Laval J.
J. Biol. Chem. 268:9063-9070(1993).

814. (G_glu_transpept) Gamma-glutamyltranspeptidase signature
30 PROSITE cross-reference(s): PS00462; G_GLU_TRANSPEPTIDASE

Gamma-glutamyltranspeptidase (EC 2.3.2.2) (GGT) [1] catalyzes the transfer of
the gamma-glutamyl moiety of glutathione to an acceptor that may be an amino
acid, a peptide or water (forming glutamate). GGT plays a key role in the

gamma-glutamyl cycle, a pathway for the synthesis and degradation of glutathione. In prokaryotes and eukaryotes, it is an enzyme that consists of two polypeptide chains, a heavy and a light subunit, processed from a single chain precursor. The active site of GGT is known to be located in the light

5   subunit.

The sequences of mammalian and bacterial GGT show a number of regions of high similarity [2]. Pseudomonas cephalosporin acylases (EC 3.5.1.-) that convert 7-beta-(4-carboxybutanamido)-cephalosporanic acid (GL-7ACA) into

10   7-aminocephalosporanic acid (7ACA) and glutaric acid are evolutionary related to GGT and also show some GGT activity [3]. Like GGT, these GL-7ACA acylases, are also composed of two subunits.

One of the conserved regions correspond to the N-terminal extremity of the

15   mature light chains of these enzymes. This region was used as a signature pattern.

Consensus patternT-[STA]-H-x-[ST]-[LIVMA]-x(4)-G-[SN]-x-V-[STA]-x-T-x-T-[LIVM]-[NE]-x(1,2)-[FY]-G

20

[1]

Tate S.S., Meister A.

Meth. Enzymol. 113:400-419(1985).

[2]

25   Suzuki H., Kumagai H., Echigo T., Tochikura T.

J. Bacteriol. 171:5169-5172(1989).

[3]

Ishiye M., Niwa M.

Biochim. Biophys. Acta 1132:233-239(1992).

30

815. G-protein gamma subunit profile

PROSITE cross-reference(s): PS50058; G_PROTEIN_GAMMA

Guanine nucleotide-binding proteins (G proteins) [1] act as intermediaries in

the transduction of signals generated by transmembrane receptors. G proteins consist of three subunits (alpha, beta, and gamma). The alpha subunit binds to and hydrolyzes GTP; the functions of the beta and gamma subunits are less clear but they seem to be required for the replacement of GDP by GTP as well

5   as for membrane anchoring and receptor recognition.

The gamma subunits are small proteins (from 70 to 110 residues) that are bound to the membrane via a isoprenyl group (either a farnesyl or a geranyl-geranyl) covalently linked to their C-terminus. In mammals there are at least

10   12 different isoforms of gamma subunits.

The Caenorhabditis elegans protein egl-10, which is a regulator of G-protein signalling, contains a G-protein gamma-like domain.

15   A profile was developed that spans the complete length of the gamma subunit.

[1]
Pennington S.R.
20   Protein Prof. 2:16-315(1995).

816. GNS1/SUR4 family signature
PROSITE cross-reference(s): PS01188; GNS1_SUR4

25   The following group of eukaryotic integral membrane proteins, whose exact function has not yet clearly been established, are evolutionary related [1]:

- Yeast GNS1 [2], a protein involved in synthesis of 1,3-beta-glucan.
- Yeast SUR4 (or APA1, SRE1) [3], a protein that could act in a glucose-
30   signaling pathway that controls the expression of several genes that are transcriptionally regulated by glucose.
- Yeast hypothetical protein YJL196c.
- Caenorhabditis elegans hypothetical protein C40H1.4.
- Caenorhabditis elegans hypothetical protein D2024.3.

The proteins have from 290 to 435 amino acid residues. Structurally, they seem

to be formed of three sections: a N-terminal region with two transmembrane

domains, a central hydrophilic loop and a C-terminal region that contains from

5   one to three transmembrane domains. A conserved region that contains three histidines was

selected as a signature pattern. This region is located in the

hydrophilic loop.

Consensus patternL-x-F-L-H-x-Y-H-H

10

[1]

Bairoch A.

Unpublished observations (1996).

[2]

15   El-Sherbeini M., Clemas J.A.

J. Bacteriol. 177:3227-3234(1995).

[3]

Garcia-Arranz M., Maldonado A.M., Mazon M.J., Portillo F.

J. Biol. Chem. 269:18076-18082(1994).

20

817. Immunoglobulins and major histocompatibility complex proteins signature

PROSITE cross-reference(s): PS00290; IG_MHC

The basic structure of immunoglobulin (Ig) [1] molecules is a tetramer of two

25   light chains and two heavy chains linked by disulfide bonds. There are two

types of light chains: kappa and lambda, each composed of a constant domain

(CL) and a variable domain (VL). There are five types of heavy chains: alpha,

delta, epsilon, gamma and mu, all consisting of a variable domain (VH) and

three (in alpha, delta and gamma) or four (in epsilon and mu) constant

30   domains (CH1 to CH4).

The major histocompatibility complex (MHC) molecules are made of two chains.

In class I [2] the alpha chain is composed of three extracellular domains, a

transmembrane region and a cytoplasmic tail. The beta chain (beta-2-

microglobulin) is composed of a single extracellular domain. In class II [3], both the alpha and the beta chains are composed of two extracellular domains, a transmembrane region and a cytoplasmic tail.

5   It is known [4,5] that the Ig constant chain domains and a single extracellular domain in each type of MHC chains are related. These homologous domains are approximately one hundred amino acids long and include a conserved intradomain disulfide bond. A small pattern around the C-terminal cysteine is involved in this disulfide bond which can be used to detect

10   these category of Ig related proteins.

Consensus pattern[FY]-x-C-x-[VA]-x-H-Sequences known to belong to this class detected by the pattern: Ig heavy chains type Alpha C region : All, in CH2 and CH3. Ig heavy chains type Delta C region : All, in CH3. Ig

15   heavy chains type Epsilon C region: All, in CH1, CH3 and CH4. Ig heavy chains type Gamma C region : All, in CH3 and also CH1 in some cases Ig heavy chains type Mu C region : All, in CH2, CH3 and CH4. Ig light chains type Kappa C region : In all CL except rabbit and Xenopus. Ig light chains type Lambda C region : In all CL except rabbit. MHC class I alpha chains :

20   All, in alpha-3 domains, including in the cytomegalovirus MHC-1 homologous protein [6]. Beta-2-microglobulin : All. MHC class II alpha chains: All, in alpha-2 domains. MHC class II beta chains: All, in beta-2 domains.

[1]

25   Gough N.

Trends Biochem. Sci. 6:203-205(1981).

[2]

Klein J., Figueroa F.

Immunol. Today 7:41-44(1986).

30   [3]

Figueroa F., Klein J.

Immunol. Today 7:78-81(1986).

[4]

Orr H.T., Lancet D., Robb R.J., Lopez de Castro J.A., Strominger J.L.

Nature 282:266-270(1979).

[5]

Cushley W., Owen M.J.

Immunol. Today 4:88-92(1983).

5 [6]

Beck S., Barrel B.G.

Nature 331:269-272(1988).

818. (IGFBP) Insulin-like growth factor binding proteins signature

10 PROSITE cross-reference(s): PS00222; IGF_BINDING

The insulin-like growth factors (IGF-I and IGF-II) bind to specific binding proteins in extracellular fluids with high affinity [1,2,3]. These IGF-binding proteins (IGFBP) prolong the half-life of the IGFs and have been shown to

15 either inhibit or stimulate the growth promoting effects of the IGFs on cells culture. They seem to alter the interaction of IGFs with their cell surface receptors. There are at least six different IGFBPs and they are structurally related.

20 The following growth-factor inducible proteins are structurally related to IGFBPs and could function as growth-factor binding proteins [4,5]:

- Mouse protein cyr61 and its probable chicken homolog, protein CEF-10.
- Human connective tissue growth factor (CTGF) and its mouse homolog, protein

25 FISP-12.

- Vertebrate protein NOV.

As a signature pattern a conserved cysteine-rich region locatedin the N-terminal section of these proteins is used.

30

Consensus patternG-C-[GS]-C-C-x(2)-C-A-x(6)-C

Sequences known to belong to this class detected by the patternALL, except for IGFBP-6's.

[1]

Rechler M.M.

Vitam. Horm. 47:1-114(1993).

[2]

5   Shimasaki S., Ling N.

Prog. Growth Factor Res. 3:243-266(1991).

[3]

Clemmons D.R.

Trends Endocrinol. Metab. 1:412-417(1990).

10   [4]

Bradham D.M., Igarashi A., Potter R.L., Grotendorst G.R.

J. Cell Biol. 114:1285-1294(1991).

[5]

Maloisel V., Martinerie C., Dambrine G., Plassiart G., Brisac M., Crochet

15   J., Perbal B.

Mol. Cell. Biol. 12:10-21(1992).

819. LMWPc : Low molecular weight phosphotyrosine protein phosphatase
Number of members: 34

20

[1]Medline: 94329182, The crystal structure of a low-molecular-weight phosphotyrosine
protein phosphatase. Su XD, Taddei N, Stefani M, Ramponi G, Nordlund P; Nature
1994;370:575-578.

25   820. (myosin_head) ATP/GTP-binding site motif A (P-loop)
      PROSITE cross-reference(s): PS00017; ATP_GTP_A

From sequence comparisons and crystallographic data analysis it has been shown
[1,2,3,4,5,6] that an appreciable proportion of proteins that bind ATP or GTP
30   share a number of more or less conserved sequence motifs. The best conserved
of these motifs is a glycine-rich region, which typically forms a flexible
loop between a beta-strand and an alpha-helix. This loop interacts with one of
the phosphate groups of the nucleotide. This sequence motif is generally
referred to as the 'A' consensus sequence [1] or the 'P-loop' [5].

There are numerous ATP- or GTP-binding proteins in which the P-loop is found. A number of protein families for which the relevance of the presence of such motif has been noted is listed below:

5

- ATP synthase alpha and beta subunits (see <PDOC00137>).
- Myosin heavy chains.
- Kinesin heavy chains and kinesin-like proteins (see <PDOC00343>).
- Dynamins and dynamin-like proteins (see <PDOC00362>).

10
- Guanylate kinase (see <PDOC00670>).
- Thymidine kinase (see <PDOC00524>).
- Thymidylate kinase (see <PDOC01034>).
- Shikimate kinase (see <PDOC00868>).
- Nitrogenase iron protein family (nifH/frxC) (see <PDOC00580>).

15
- ATP-binding proteins involved in 'active transport' (ABC transporters) [7] (see <PDOC00185>).
- DNA and RNA helicases [8,9,10].
- GTP-binding elongation factors (EF-Tu, EF-1alpha, EF-G, EF-2, etc.).
- Ras family of GTP-binding proteins (Ras, Rho, Rab, Ral, Ypt1, SEC4, etc.).

20
- Nuclear protein ran (see <PDOC00859>).
- ADP-ribosylation factors family (see <PDOC00781>).
- Bacterial dnaA protein (see <PDOC00771>).
- Bacterial recA protein (see <PDOC00131>).
- Bacterial recF protein (see <PDOC00539>).

25
- Guanine nucleotide-binding proteins alpha subunits (Gi, Gs, Gt, G0, etc.).
- DNA mismatch repair proteins mutS family (See <PDOC00388>).
- Bacterial type II secretion system protein E (see <PDOC00567>).

Not all ATP- or GTP-binding proteins are picked-up by this motif. A number of

30
proteins escape detection because the structure of their ATP-binding site is completely different from that of the P-loop. Examples of such proteins are the E1-E2 ATPases or the glycolytic kinases. In other ATP- or GTP-binding proteins the flexible loop exists in a slightly different form; this is the case for tubulins or protein kinases. A special mention must be reserved for

adenylate kinase, in which there is a single deviation from the P-loop
pattern: in the last position Gly is found instead of Ser or Thr.

Consensus pattern[AG]-x(4)-G-K-[ST]

5

[1]

Walker J.E., Saraste M., Runswick M.J., Gay N.J.

EMBO J. 1:945-951(1982).

[2]

10   Moller W., Amons R.

FEBS Lett. 186:1-7(1985).

[3]

Fry D.C., Kuby S.A., Mildvan A.S.

Proc. Natl. Acad. Sci. U.S.A. 83:907-911(1986).

15   [4]

Dever T.E., Glynias M.J., Merrick W.C.

Proc. Natl. Acad. Sci. U.S.A. 84:1814-1818(1987).

[5]

Saraste M., Sibbald P.R., Wittinghofer A.

20   Trends Biochem. Sci. 15:430-434(1990).

[6]

Koonin E.V.

J. Mol. Biol. 229:1165-1174(1993).

[7]

25   Higgins C.F., Hyde S.C., Mimmack M.M., Gileadi U., Gill D.R., Gallagher
M.P.

J. Bioenerg. Biomembr. 22:571-592(1990).

[8]

Hodgman T.C.

30   Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).

[9]

Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K.,
Schnier J., Slonimski P.P.

Nature 337:121-122(1989).

[10]

Gorbalenya A.E., Koonin E.V., Donchenko A.P., Blinov V.M.

Nucleic Acids Res. 17:4713-4730(1989).

5   821. PE: PE family

This family named after a PE motif near to the amino terminus of the domain. The PE family

of proteins all contain an amino-terminal region of about 110 amino acids. The carboxyl

terminus of this family are variable and fall into several classes. The largest class of PE

proteins is the highly repetitive PGRS class which have a high glycine content. The function

10   of these proteins is uncertain but it has been suggested that they may be related to antigenic

variation of Mycobacterium tuberculosis [1]. Number of members: 88

[1] Medline: 98295987. Deciphering the biology of Mycobacterium tuberculosis from the

complete genome sequence. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D,

15   Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D,

Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd

S, Hornsby T, Jagels K, Barrell BG, et al; Nature 1998;393:537-544.

822. (RNB) Ribonuclease II family signature

20   PROSITE cross-reference(s): PS01175; RIBONUCLEASE_II

On the basis of sequence similarities, the following bacterial and eukaryotic

proteins seem to form a family:

25   - Escherichia coli and related bacteria ribonuclease II (EC 3.1.13.1) (RNase

II) (gene rnb) [1]. RNase II is an exonuclease involved in mRNA decay. It

degrades mRNA by hydrolyzing single-stranded polyribonucleotides

processively in the 3' to 5' direction.

- Bacterial protein vacB. In Shigella flexneri, vacB has been shown to be

30   required for the expression of virulence genes at the posttranscriptional

level.

- Yeast protein SSD1 (or SRK1) which is implicated in the control of the cell

cycle G1 phase.

- Yeast protein DIS3 [2], which binds to ran (GSP1) and ehances the the

nucleotide-releasing activity of RCC1 on ran.

- Fission yeast protein dis3, which is implicated in mitotic control.

- Neurospora crassa cyt-4, a mitochondrial protein required for RNA 5' and 3' end processing and splicing.

5 - Yeast protein MSU1, which is involved in mitochondrial biogenesis.

- Synechocystis strain PCC 6803 protein zam [3], which control resistance to the carbonic anhydrase inhibitor acetazolamide.

- Caenorhabditis elegans hypothetical protein F48E8.6.

10 The size of these proteins range from 644 residues (rnb) to 1250 (SSD1). While their sequence is highly divergent they share a conserved domain in their C-terminal section [4]. It is possible that this domain plays a role in a putative exonuclease function that would be common to all these proteins. A signature pattern was developed based on the core of this conserved domain.

15

Consensus pattern[HI]-[FYE]-[GSTAM]-[LIVM]-x(4,5)-Y-[STAL]-x-[FWVAC]-[TV]-[SA]-P-[LIVMA]-[RQ]-[KR]-[FY]-x-D-x(3)-[HQ]

[1]

20 Zilhao R., Camelo L., Arraiano C.M.

Mol. Microbiol. 8:43-51(1993).

[2]

Noguchi E., Hayashi N., Azuma Y., Seki T., Nakamura M., Nakashima N., Yanagida M., He X., Mueller U., Sazer S., Nishimoto T.

25 EMBO J. 15:5595-5605(1996).

[3]

Beuf L., Bedu S., Cami B., Joset F.

Plant Mol. Biol. 27:779-788(1995).

[4]

30 Mian I.S.

Nucleic Acids Res. 25:3187-3195(1997).

823. Src homology 2 (SH2) domain profile
PROSITE cross-reference(s): PS50001; SH2

The Src homology 2 (SH2) domain is a protein domain of about 100 amino-acid residues first identified as a conserved sequence region between the oncoproteins Src and Fps [1]. Similar sequences were later found in many other

5   intracellular signal-transducing proteins [2]. SH2 domains function as regulatory modules of intracellular signalling cascades by interacting with high affinity to phosphotyrosine-containing target peptides in a sequence-specific and strictly phosphorylation-dependent manner [3,4,5,6].

10   The SH2 domain has a conserved 3D structure consisting of two alpha helices and six to seven beta-strands. The core of the domain is formed by a continuous beta-meander composed of two connected beta-sheets [7].

So far, SH2 domains have been identified in the following proteins:

15

- Many vertebrate, invertebrate and retroviral cytoplasmic (non-receptor) protein tyrosine kinases. In particular in the Src, Abl, Bkt, Csk and ZAP70 families of kinases.
- Mammalian phosphatidylinositol-specific phospholipase C gamma-1 and -2. Two

20   copies of the SH2 domain are found in those proteins in between the catalytic 'X-' and 'Y-boxes' (see <PDOC50007>).
- Mammalian phosphatidyl inositol 3-kinase regulatory p85 subunit.
- Some vertebrate and invertebrate protein-tyrosine phosphatases.
- Mammalian Ras GTPase-activating protein (GAP).

25   - Adaptor proteins mediating binding of guanine nucleotide exchange factors to growth factor receptors: vertebrate GRB2, Caenorhabditis elegans sem-5 and Drosophila DRK.
- Mammalian Vav oncoprotein, a guanine-nucleotide exchange factor of the CDC24 family.

30   - Miscellanous proteins interacting with vertebrate receptor protein tyrosine kinases: oncoprotein Crk, mammalian cytoplasmic proteins Nck, Shc.
- STAT proteins (signal transducers and activators of transcription).
- Chicken tensin.
- Yeast transcriptional control protein SPT6.

The profile developed to detect SH2 domains is based on a structural alignment consisting of 8 gap-free blocks and 7 linker regions totaling 92 match positions.

5

[1]

Sadowski I., Stone J.C., Pawson T.

Mol. Cell. Biol. 6:4396-4408(1986).

[2]

10   Russel R.B., Breed J., Barton G.J.

FEBS Lett. 304:15-20(1992).

[3]

Marangere L.E.M., Pawson T.

J. Cell Sci. Suppl. 18:97-104(1994).

15   [4]

Pawson T., Schlessinger J.

Curr. Biol. 3:434-442(1993).

[5]

Mayer B.J., Baltimore D.

20   Trends Cell. Biol. 3:8-13(1993).

[6]

Pawson T.

Nature 373:573-580(1995).

[7]

25   Kuriyan J., Cowburn D.

Curr. Opin. Struct. Biol. 3:828-837(1993).

824. Sulfate transporters signature

PROSITE cross-reference(s): PS01130; SULFATE_TRANSP

30

A number of proteins involved in the transport of sulfate across a membrane as well as some yet uncharacterized proteins have been shown [1,2] to be evolutionary related. These proteins are:

- Neurospora crassa sulfate permease II (gene cys-14).

- Yeast sulfate permeases (genes SUL1 and SUL2).

- Rat sulfate anion transporter 1 (SAT-1).

- Mammalian DTDST, a probable sulfate transporter which, in Human, is

5   involved in the genetic disease, diastrophic dysplasia (DTD).

- Sulfate transporters 1, 2 and 3 from the legume Stylosanthes hamata.

- Human pendrin (gene PDS), which is involved in a number of hearing loss

genetic diseases.

10   - Human protein DRA (Down-Regulated in Adenoma).

- Soybean early nodulin 70.

- Escherichia coli hypothetical protein ychM.

- Caenorhabditis elegans hypothetical protein F41D9.5.

15   As expected by their transport function, these proteins are highly hydrophobic

and seem to contain about 12 transmembrane domains. The best conserved region

seems to be located in the second transmembrane region and is used as a

signature pattern.

20   Consensus pattern[PAV]-x-Y-[GS]-L-Y-[STAG](2)-x(4)-[LIVFYA]-[LIVST]-[YI]-
x(3)-[GA]-[GST]-S-[KR]

[1]

Sandal N.N., Marcker K.A.

25   Trends Biochem. Sci. 19:19-19(1994).

[2]

Smith F.W., Hawkesford M.J., Prosser I.M., Clarkson D.T.

Mol. Gen. Genet. 247:709-715(1995).

30   825. TYA: TYA transposon protein

Ty are yeast transposons. A 5.7kb transcript codes for p3 a fusion protein of TYA and TYB.

The TYA protein is analogous to the gag protein of retroviruses. TYA a is cleaved to form

46kd protein which can form mature virion like particles [1]. Number of members: 59

[1] Medline: 97404699. Cryo-electron microscopy structure of yeast Ty retrotransposon virus-like particles. Palmer KJ, Tichelaar W, Myers N, Burns NR, Butcher SJ, Kingsman AJ, Fuller SD, Saibil HR; J Virol 1997;71:6863-6868.

826. Aldolase_II

Class II Aldolase and Adducin N-terminal domain.

-!- This family includes class II aldolases and adducins which have not been ascribed any enzymatic function. Number of members: 37

References:

[1] Medline: 93294819. The spatial structure of the class II L-fuculose-1-phosphate aldolase from Escherichia coli. Dreyer MK, Schulz GE; J Mol Biol 1993;231:549-553.

[2] Medline: 96256522. Catalytic mechanism of the metal-dependent fuculose aldolase from Escherichia coli as derived from the structure. Dreyer MK, Schulz GE; J Mol Biol 1996;259:458-466.

827. CBD_2

-!- Two tryptophan residues are involved in cellulose binding.

-!- Cellulose binding domain found in bacteria. Number of members: 51

References:

[1] Medline: 95284032. Solution structure of a cellulose-binding domain from Cellulomonas fimi by nuclear magnetic resonance spectroscopy. Xu GY, Ong E, Gilkes NR, Kilburn DG, Muhandiram DR, Harris-Brandts M, Carver JP, Kay LE, Harvey TS; Biochemistry 1995;34:6993-7009.

828. P

A unique feature of the eukaryotic subtilisin-like proprotein convertases is the presence of an additional highly conserved sequence of approximately 150 residues (P domain) located immediately downstream of the catalytic domain.
Number of members: 91

References:

[1] Medline: 94252314. A C-terminal domain conserved in precursor processing proteases is required for intramolecular N-terminal maturation of pro-Kex2 protease. Gluschankof P, Fuller RS; EMBO J 1994;13:2280-2288.

[2] Medline: 98225190. Regulatory roles of the P domain of the subtilisin-like prohormone convertases. Zhou A, Martin S, Lipkind G, LaMendola J, Steiner DF; J Biol Chem 1998;273:11107-11114.

829. Uncharacterized protein family UPF0020 signature
PROSITE cross-reference(s): PS01261; UPF0020
The following uncharacterized proteins have been shown [1] to share regions of similarities:

- Escherichia coli hypothetical protein ycbY and HI0116/15, the corresponding Haemophilus influenzae protein.
- Bacillus subtilis hypothetical protein ypsC.
- Synechocystis strain PCC 6803 hypothetical protein slr0064.
- Methanococcus jannaschii hypothetical proteins MJ0438 and MJ0710.

These are hydrophilic proteins of from 40 Kd to about 80 Kd. They can be picked up in the database by the following pattern.

Consensus patternD-P-[LIVMF]-C-G-[ST]-G-x(3)-[LI]-E

References:

[1] Bairoch A. Unpublished observations (1997).

830. Uncharacterized protein family UPF0031 signatures
PROSITE cross-reference(s): PS01049; UPF0031_1; PS01050; UPF0031_2
The following uncharacterized proteins have been shown [1] to share regions of similarities:

- Yeast chromosome XI hypothetical protein YKL151c.
- Caenorhabditis elegans hypothetical protein R107.2.
- Escherichia coli hypothetical protein yjeF.

- Bacillus subtilis hypothetical protein yxkO.

- Helicobacter pylori hypothetical protein HP1363.

- Mycobacterium tuberculosis hypothetical protein MtCY77.05c.

- Mycobacterium leprae hypothetical protein B229_C2_201.

5   - Synechocystis strain PCC 6803 hypothetical protein sll1433.

- Methanococcus jannaschii hypothetical protein MJ1586.

These are proteins of about 30 to 40 Kd whose central region is well conserved. They can be picked up in the database by the following patterns.

10

Consensus pattern[SAV]-[IVW]-[LVA]-[LIV]-G-[PNS]-G-L-[GP]-x-[DENQT]
Consensus pattern[GA]-G-x-G-D-[TV]-[LT]-[STA]-G-x-[LIVM]

831. (ACOX)

15   Acyl-CoA oxidase

This is a family of Acyl-CoA oxidases EC:1.3.3.6. Acyl-coA oxidase converts acyl-CoA into trans-2-enoyl-CoA [1].

20   Number of members: 39

[1] Hayashi H, De Bellis L, Yamaguchi K, Kato A, Hayashi M, Nishimura M; Medline: 98192624. Molecular characterization of a glyoxysomal long chain acyl-CoA oxidase that is synthesized as a precursor of higher molecular mass in pumpkin." J Biol Chem
25   1998;273:8301-8307.

832. (AICARFT_IMPCHas)
AICARFT/IMPCHase bienzyme

30

This is a family of bifunctional enzymes catalysing the last steps in de novo purine biosynthesis. The bifunctional enzyme is found in both prokaryotes and eukaryotes. The second last step is catalysed by 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase EC:2.1.2.3 (AICARFT), this enzyme catalyses the formylation of AICAR

with 10-formyl-tetrahydrofolate to yield FAICAR and tetrahydrofolate [1]. The last step is catalysed by IMP (Inosine monophosphate) cyclohydrolase EC:3.5.4.10 (IMPCHase), cyclizing FAICAR (5-formylaminoimidazole-4-carboxamide ribonucleotide) to IMP [1].

5   Number of members: 22

[1] Akira T, Komatsu M, Nango R, Tomooka A, Konaka K, Yamauchi M, Kitamura Y, Nomura S, Tsukamoto I; Medline: 97473523   Molecular cloning and expression of a rat cDNA encoding 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP
10   cyclohydrolase" [published erratum appears in Gene 1998 Feb 27;208(2):337] Gene 1997;197:289-293.

[2] Rayl EA, Moroson BA, Beardsley GP; Medline: 96147205   The human purH gene product, 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase. Cloning, sequencing, expression, purification, kinetic analysis, and domain
15   mapping." J Biol Chem 1996;271:2225-2233.

833. (AOX)

Alternative oxidase

20

The alternative oxidase is used as a second terminal oxidase in the mitochondria, electrons are transfered directly from reduced ubiquinol to oxygen forming water [2]. This is not coupled to ATP synthesis and is not inhibited by cyanide, this pathway is a single step process [1]. In rice the transcript levels of the alternative oxidase are increased by low
25   temperature [1].

Number of members: 27

[1] Ito Y, Saisho D, Nakazono M, Tsutsumi N, Hirai A; Medline: 98086211   Transcript
30   levels of tandem-arranged alternative oxidase genes in rice are increased by low temperature." Gene 1997;203:121-129.

●

667

[2] Li Q, Ritzel RG, McLean LL, McIntosh L, Ko T, Bertrand H, Nargang FE; Medline: 96366413 Cloning and analysis of the alternative oxidase gene of Neurospora crassa." Genetics 1996;142:129-140.

5

834. (APH)

Protein kinases signatures and profile

Cross-reference(s): PS00107; PROTEIN_KINASE_ATP, PS00108;
10 PROTEIN_KINASE_ST, PS00109; PROTEIN_KINASE_TYR, PS50011;
PROTEIN_KINASE_DOM

Eukaryotic protein kinases [1 to 5] are enzymes that belong to a very extensive family of proteins which share a conserved catalytic core common to both serine/threonine and tyrosine
15 protein kinases. There are a number of conserved regions in the catalytic domain of protein kinases. Two of these regions have been selected to build signature patterns. The first region, which is located in the N-terminal extremity of the catalytic domain, is a glycine-rich stretch of residues in the vicinity of a lysine residue, which has been shown to be involved in ATP binding. The second region, which is located in the central part of the catalytic domain,
20 contains a conserved aspartic acid residue which is important for the catalytic activity of the enzyme [6]; two signature patterns were derived for that region: one specific for serine/ threonine kinases and the other for tyrosine kinases. A profile was developed which is based on the alignment in [1] and covers the entire catalytic domain.

25 Consensus pattern: [LIV]-G-{P}-G-{P}-[FYWMGSTNH]-[SGA]-{PW}-[LIVCAT]-{PD}-x-[GSTACLIVMFY]-x(5,18)-[LIVMFYWCSTAR]-[AIVP]-[LIVMFAGCKR]-K [K binds ATP]

Sequences known to belong to this class detected by the pattern the majority of known
30 protein kinases but it fails to find a number of them, especially viral kinases which are quite divergent in this region and are completely missed by this pattern.

Consensus pattern: [LIVMFYC]-x-[HY]-x-D-[LIVMFY]-K-x(2)-N-[LIVMFYCT](3) [D is an active site residue]

Sequences known to belong to this class detected by the pattern. Most serine/ threonine specific protein kinases with 10 exceptions (half of them viral kinases) and also Epstein-Barr virus BGLF4 and Drosophila ninaC which have respectively Ser and Arg instead of the

5   conserved Lys and which are therefore detected by the tyrosine kinase specific pattern described below.

Consensus pattern: [LIVMFYC]-x-[HY]-x-D-[LIVMFY]-[RSTAC]-x(2)-N-[LIVMFYC](3) [D is an active site residue] tyrosine specific protein kinases with the exception of human

10 ERBB3 and mouse blk. This pattern will also detect most bacterial aminoglycoside phosphotransferases [8,9] and herpesviruses ganciclovir kinases [10]; which are proteins structurally and evolutionary related to protein kinases. Sequences known to belong to this class detected by the profile ALL, except for three viral kinases. This profile also detects receptor guanylate cyclases (see <PDOC00430>) and 2-5A-dependent ribonucleases.

15 Sequence similarities between these two families and the eukaryotic protein kinase family have been noticed before. It also detects Arabidopsis thaliana kinase- like protein TMKL1 which seems to have lost its catalytic activity.

Note if a protein analyzed includes the two protein kinase signatures, the probability of it

20 being a protein kinase is close to 100%. Note eukaryotic-type protein kinases have also been found in prokaryotes such as Myxococcus xanthus [11] and Yersinia pseudotuberculosis. Note the patterns shown above has been updated since their publication in [7]. Note this documentation entry is linked to both signature patterns and a profile. As the profile is much more sensitive than the patterns, you should use it if you have access to the necessary

25 software tools to do so.

References

[1] Hanks S.K., Hunter T., FASEB J. 9:576-596(1995).

[2] Hunter T., Meth. Enzymol. 200:3-37(1991).

30 [3] Hanks S.K., Quinn A.M., Meth. Enzymol. 200:38-62(1991).

[4] Hanks S.K., Curr. Opin. Struct. Biol. 1:369-383(1991).

[5] Hanks S.K., Quinn A.M., Hunter T., Science 241:42-52(1988).

[6] Knighton D.R., Zheng J., Ten Eyck L.F., Ashford V.A., Xuong N.-H., Taylor, S.S., Sowadski J.M., Science 253:407-414(1991).

[7] Bairoch A., Claverie J.-M., Nature 331:22(1988).

[8] Benner S., Nature 329:21-21(1987).

[9] Kirby R., J. Mol. Evol. 30:489-492(1992).

[10] Littler E., Stuart A.D., Chee M.S., Nature 358:160-162(1992).

5   [11] Munoz-Dorado J., Inouye S., Inouye M., Cell 67:995-1006(1991).

835. (Asp_Glu_race)

Aspartate and glutamate racemases signatures

10

Cross-reference(s) PS00923; ASP_GLU_RACEMASE_1 PS00924;
ASP_GLU_RACEMASE_2

Aspartate racemase (EC 5.1.1.13) and glutamate racemase (EC 5.1.1.3) are two evolutionary
15   related bacterial enzymes that do not seem to require a cofactor for their activity [1].
Glutamate racemase, which interconverts L-glutamate into D-glutamate, is required for the
biosynthesis of peptidoglycan and some peptide-based antibiotics such as gramicidin S. In
addition to characterized aspartate and glutamate racemases, this family also includes a
hypothetical protein from Erwinia carotovora and one from Escherichia coli (ygeA). Two
20   conserved cysteines are present in the sequence of these enzymes. They are expected to play
a role in catalytic activity by acting as bases in proton abstraction from the substrate.
Signature patterns were developed for both cysteines.

Consensus pattern: [IVA]-[LIVM]-x-C-x(0,1)-N-[ST]-[MSA]-[STH]-[LIVFYSTANK]
25

Consensus pattern: [LIVM](2)-x-[AG]-C-T-[DEH]-[LIVMFY]-[PNGRS]-x-[LIVM]

[1] Gallo K.A., Knowles J.R., Biochemistry 32:3981-3990(1993).

30

836. (ATP-sulfurylase)

ATP-sulfurylase

670

This family consists of ATP-sulfurylase or sulfate adenylyltransferase EC:2.7.7.4 some of which are part of a bifunctional polypeptide chain associated with adenosyl phosphosulphate (APS) kinase APS_kinase. Both enzymes are required for PAPS (phosphoadenosine-phosphosulfate) synthesis from inorganic sulphate [2]. ATP sulfurylase catalyses the synthesis of adenosine-phosphosulfate APS from ATP and inorganic sulphate [1].

Number of members: 37

[1] Kurima K, Warman ML, Krishnan S, Domowicz M, Krueger RC Jr, Deyrup A, Schwartz NB; Medline: 98337975   A member of a family of sulfate-activating enzymes causes murine brachymorphism" [published erratum appears in Proc Natl Acad Sci U S A 1998 Sep 29;95(20):12071] Proc Natl Acad Sci U S A 1998;95:8681-8685.

[2] Rosenthal E, Leustek T; Medline: 96096529   A multifunctional Urechis caupo protein, PAPS synthetase, has both ATP sulfurylase and APS kinase activities." Gene 1995;165:243-248.

837. (ATP-synt_F)
ATP synthase (F/14-kDa) subunit

This family includes 14-kDa subunit from vATPases [1], which is in the peripheral catalytic part of the complex [2]. The family also includes archaebacterial ATP synthase subunit F [3].

Number of members: 23

[1] Guo Y, Kaiser K, Wieczorek H, Dow JA; Medline: 96269411   The Drosophila melanogaster gene vha14 encoding a 14-kDa F-subunit of the vacuolar ATPase." Gene 1996;172:239-243.

[2] Peng SB, Crider BP, Tsai SJ, Xie XS, Stone DK; Medline: 96216416   Identification of a 14-kDa subunit associated with the catalytic sector of clathrin-coated vesicle H+-ATPase." J Biol Chem 1996;271:3324-3327.

671

[3] Wilms R, Freiberg C, Wegerle E, Meier I, Mayer F, Muller V; Medline: 96324968 Subunit structure and organization of the genes of the A1A0 ATPase from the Archaeon Methanosarcina mazei Go1." J Biol Chem 1996;271:18843-18852.

838. (CBD_4)

Starch binding domain

Number of members: 48

839. (CbiX)

The function of CbiX is uncertain, however it is found in cobalamin biosynthesis operons and so may have a related function. Some CbiX proteins contain a striking histidine-rich region at their C-terminus, which suggests that it might be involved in metal chelation [1].

Number of members: 6

[1] Raux E, Lanois A, Warren MJ, Rambach A, Thermes C; Medline: 98416126   Cobalamin (vitamin B12) biosynthesis: identification and characterization of a Bacillus megaterium cobI operon." Biochem J 1998;335:159-166.

840. (Complex1_51K)

Respiratory-chain NADH dehydrogenase 51 Kd subunit signatures Cross-reference(s) PS00644; COMPLEX1_51K_1 PS00645; COMPLEX1_51K_2

Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex there is one with a molecular weight of 51 Kd (in mammals),

which is the second largest subunit of complex I and is a component of the iron-sulfur (IP) fragment of the enzyme. It seems to bind to NAD, FMN, and a 2Fe-2S cluster.

The 51 Kd subunit is highly similar to [3,4]:

5 - Subunit alpha of Alcaligenes eutrophus NAD-reducing hydrogenase (gene hoxF) which also binds to NAD, FMN, and a 2Fe-2S cluster.

- Subunit NQO1 of Paracoccus denitrificans NADH-ubiquinone oxidoreductase.

- Subunit F of Escherichia coli NADH-ubiquinone oxidoreductase (gene nuoF).

10 The 51 Kd subunit and the bacterial hydrogenase alpha subunit contains three regions of sequence similarities. The first one most probably corresponds to the NAD-binding site, the second to the FMN-binding site, and the third one, which contains three cysteines, to the iron-sulfur binding region. Signature patterns have been developed for the FMN-binding and for the 2Fe-2S binding regions.

15

Consensus pattern: G-[AM]-G-[AR]-Y-[LIVM]-C-G-[DE](2)-[STA](2)-[LIM](2)-[EN]- S

Consensus pattern: E-S-C-G-x-C-x-P-C-R-x-G [The three C's are putative 2Fe-2S ligands]

[1] Ragan C.I., Curr. Top. Bioenerg. 15:1-36(1987).

20 [2] Weiss H., Friedrich T., Hofhaus G., Preis D., Eur. J. Biochem. 197:563-576(1991).

[3] Fearnley I.M., Walker J.E. Biochim. Biophys. Acta 1140:105-134(1992).

[4] Weidner U., Geier S., Ptock A., Friedrich T., Leif H., Weiss H., J. Mol. Biol. 233:109-122(1993).

25

841. (DAP_epimerase)

Diaminopimelate epimerase signature

Cross-reference(s) PS01326; DAP_EPIMERASE

30 Diaminopimelate epimerase (EC 5.1.1.7) catalyzes the isomeriazation of L,L- to D,L-meso-diaminopimelate in the biosynthetic pathway leading from aspartate to lysine. This enzyme is a protein of about 30 Kd. Two conserved cysteines seem [1] to function as the acid and base in the catalytic mechanism. As a signature pattern, the region surrounding the first of these two active site cysteines were selected.

Consensus pattern: N-x-D-G-S-x(4)-C-G-N-[GA]-x-R [C is an active site residue] Sequences known to belong to this class detected by the pattern ALL, except for an Anabaena dapF which has a Ser instead of the active site Cys.

5

[1] Cirilli M., Zheng R., Scapin G., Blanchard J.S., Biochemistry 37:16452-16458(1998).

842. (DNA_gyraseB_C)

10   DNA topoisomerase II signature

Cross-reference(s) PS00177; TOPOISOMERASE_II

DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type II topoisomerases are ATP-
15   dependent and act by passing a DNA segment through a transient double-strand break. Topoisomerase II is found in phages, archaebacteria, prokaryotes, eukaryotes, and in African Swine Fever virus (ASF). In bacteriophage T4 topoisomerase II consists of three subunits (the product of genes 39, 52 and 60). In prokaryotes and in archaebacteria the enzyme, known as DNA gyrase, consists of two subunits (genes gyrA and gyrB [E2]). In
20   some bacteria, a second type II topoisomerase has been identified; it is known as topoisomerase IV and is required for chromosome segregation, it also consists of two subunits (genes parC and parE). In eukaryotes, type II topoisomerase is a homodimer.

There are many regions of sequence homology between the different subtypes of
25   topoisomerase II. The relation between the different subunits is shown in the following representation:

```
<---------------About-1400-residues--------------------->

30    [----------Protein 39-*-----][----Protein 52----]        Phage T4
      [----------gyrB-------*-----][--------gyrA----------------]  Prokaryote II
                                   Archaebacteria
      [----------parE-------*-----][--------parD----------------]  Prokaryote IV
      [--------------------*----------------------------------]  Eukaryote and
```

'*': Position of the pattern.

As a signature pattern for this family of proteins, a region that contains a highly conserved
pentapeptide was selected. The pattern is located in gyrB, in parE, and in protein 39 of phage
T4 topoisomerase.

Consensus pattern: [LIVMA]-x-E-G-[DN]-S-A-x-[STAG]

[1] Sternglanz R., Curr. Opin. Cell Biol. 1:533-535(1990).

[2] Bjornsti M.-A., Curr. Opin. Struct. Biol. 1:99-103(1991).

[3] Sharma A., Mondragon A., Curr. Opin. Struct. Biol. 5:39-47(1995).

[4] Roca J., Trends Biochem. Sci. 20:156-160(1995).

843. (DUF16)

Protein of unknown function

The function of this protein is unknown. It appears to only occur in Mycoplasma
pneumoniae.

Number of members: 26

[1] Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R; Medline: 97105885
Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae."
Nucleic Acids Res 1996;24:4420-4449.

844. (DUF21)

Domain of unknown function

This transmembrane region has no known function. Many of the sequences in this family are
annotated as hemolysins, however this is due to a similarity to Swiss:Q54318 that does not

contain this domain. This domain is found in the N-terminus of the proteins adjacent to two intracellular CBS domains CBS.

Number of members: 42

845. (DUF56)

Integral membrane protein

The members of this family are putative integral membrane proteins. The function of the family is unknown, however the family includes Sec59 from yeast. Sec59 is a dolichol kinase EC:2.7.1.108, but it is not clear if the enzymatic activity resides in this region or its N terminal region.

Number of members: 13

846. (DUF94)

Domain of unknown function

The function of this domain is unknown. It is found in both eukaryotes and archaebacteria. The alignment contains a completely conserved aspartate residue that may be functionally important. The eukaryotic domains contains three conserved cysteines and a histidine that might be metal binding, however these are absent in the archaebacterial proteins.

Number of members: 9

847. (FF)

FF domain

This domain may be involved in protein-protein interaction [1].

Number of members: 42

5   [1] Bedford MT, Leder P; Medline: 99322199   The FF domain: a novel motif that often accompanies WW domains." Trends Biochem Sci 1999;24:264-265.

848. (FLO_LFY)
10   Floricaula / Leafy protein

This family consists of various plant development proteins which are homologues of floricaula (FLO) and Leafy (LFY) proteins which are floral meristem identity proteins. Mutations in the sequences of these proteins affect flower and leaf development.

15   Number of members: 16

[1] Hofer J, Turner L, Hellens R, Ambrose M, Matthews P, Michael A, Ellis N; Medline: 97411151   UNIFOLIATA regulates leaf and flower morphogenesis in pea." Curr Biol
20   1997;7:581-587.

[2] Weigel D, Alvarez J, Smyth DR, Yanofsky MF, Meyerowitz EM; Medline: 92274452 LEAFY controls floral meristem identity in Arabidopsis." Cell 1992;69:843-859.

25   849. (G-patch)
G-patch domain

This domain is found in a number of RNA binding proteins, and is also found in proteins that contain RNA binding domains. This suggests that this domain may have an RNA binding
30   function. This domain has seven highly conserved glycines.

Number of members: 47

[1] Aravind L, Koonin EV; Medline: 10470032 G-patch: a new conserved domain in eukaryotic RNA-processing proteins and type D retroviral polyproteins." Trends Biochem Sci 1999;24:342-344.

5

850. (Gram-ve_porins)

General diffusion Gram-negative porins signature

Cross-reference(s) PS00576; GRAM_NEG_PORIN

10 The outer membrane of Gram-negative bacteria acts as a molecular filter for hydrophilic compounds. Proteins, known as porins [1], are responsible for the 'molecular sieve' properties of the outer membrane. Porins form large water- filled channels which allows the diffusion of hydrophilic molecules into the periplasmic space. Some porins form general diffusion channels that allows any solutes up to a certain size (that size is known as the exclusion limit)

15 to cross the membrane, while other porins are specific for a solute and contain a binding site for that solute inside the pores (these are known as selective porins). As porins are the major outer membrane proteins, they also serve as receptor sites for the binding of phages and bacteriocins. General diffusion porins generally assemble as trimer in the membrane and the transmembrane core of these proteins is composed exclusively of beta strands [2]. It has been

20 shown [3] that a number of general porins are evolutionary related, these porins are:

- Enterobacteria phoE.

- Enterobacteria ompC.

- Enterobacteria ompF.

- Enterobacteria nmpC.

25 - Bacteriophage PA-2 LC.

- Neisseria PI.A.

- Neisseria PI.B.

As a signature pattern a conserved region was selected, located in the C-terminal part of these

30 proteins, which spans two putative transmembrane beta strands.

Consensus pattern: [LIVMFY]-x(2)-G-x(2)-Y-x-F-x-K-x(2)-[SN]-[STAV]-[LIVMFYW]- V

[1] Benz R., Bauer K., Eur. J. Biochem. 176:1-19(1988).

[2] Jap B.K., Walian P.J., Q. Rev. Biophys. 23:367-403(1990).

[3] Jeanteur D., Lakey J.H., Pattus F., Mol. Microbiol. 5:2153-2164(1991).

5   851. (HlyD)

HlyD family secretion proteins signature

Cross-reference(s) PS00543; HLYD_FAMILY

Gram-negative bacteria produce a number of proteins which are secreted into the growth

10   medium by a mechanism that does not require a cleaved N-terminal signal sequence. These

proteins, while having different functions, require the help of two or more proteins for their

secretion across the cell envelope. Amongst which a protein belonging to the ABC

transporters family (see the relevant entry <PDOC00185>) and a protein belonging to a

family which is currently composed [1 to 5] of the following members:

15   Gene Species   Protein which is exported

     ---- --------------------- ---

     hlyD Escherichia coli   Hemolysin

     appD A.pleuropneumoniae   Hemolysin

     lcnD Lactococcus lactis   Lactococcin A

20   lktD A.actinomycetemcomitans Leukotoxin

           Pasteurella haemolytica

     rtxD A.pleuropneumoniae   Toxin-III

     cyaD Bordetella pertussis Calmodulin-sensitive adenylate cyclase-

                                 hemolysin (cyclolysin)

25   cvaA Escherichia coli   Colicin V

     prtE Erwinia chrysanthemi Extracellular proteases B and C

     aprE Pseudomonas aeruginosa Alkaline protease

     emrA Escherichia coli   Drugs and toxins

     yjcR Escherichia coli   Unknown

30   These proteins are evolutionary related and consist of from 390 to 480 amino acid residues.

They seem to be anchored in the inner membrane by a N-terminal transmembrane region.

Their exact role in the secretion process is not yet known. The C-terminal section of these

proteins is the best conserved region; a signature pattern from that region was derived.

Consensus pattern: [LIVM]-x(2)-G-[LM]-x(3)-[STGAV]-x-[LIVMT]-x-[LIVMT]-[GE]-x-[KR]-x-[LIVMFYW](2)-x-[LIVMFYW](3)

Sequences known to belong to this class detected by the pattern ALL, except for emrA and yjcR.

References:

[1] Gilson L., Mahanty H.K., Kolter R., EMBO J. 9:3875-3884(1990).

[2] Letoffe S., Delepelaire P., Wandersman C., EMBO J. 9:1375-1382(1990).

[3] Stoddard G.W., Petzel J.P., van Belkum M.J., Kok J., McKay L.L., Appl. Environ. Microbiol. 58:1952-1961(1992).

[4] Duong F., Lazdunski A., Cami B., Murgier M., Gene 121:47-54(1992).

[5] Lewis K., Trends Biochem. Sci. 19:119-123(1994).

852. (IBR)

In Between Ring fingers

The IBR (In Between Ring fingers) domain is found to occur between pairs of ring fingers (zf-C3HC4). The function of this domain is unknown. This domain has also been called the C6HC domain and DRIL (for double RING finger linked) domain [2].

Number of members: 25

[1] Morett E, Bork P; Medline: 10366851   A novel transactivation domain in parkin."Trends Biochem Sci 1999;24:229-231.

[2] van der Reijden BA, Erpelinck-Verschueren CA, Lowenberg B, Jansen JH; Medline: 99349709   TRIADs: a new class of proteins with a novel cysteine-rich signature." Protein Sci 1999;8:1557-1561.

853. (IPPT)

IPP transferase

680

[1] Durand JM, Bjork GR, Kuwae A, Yoshikawa M, Sasakawa C; Medline: 97440126 The modified nucleoside 2-methylthio-N6-isopentenyladenosine in tRNA of Shigella flexneri is required for expression of virulence genes." J Bacteriol 1997;179:5777-5782.

5   [2] Boguta M, Hunter LA, Shen WC, Gillman EC, Martin NC, Hopper AK; Medline: 94187700 Subcellular locations of MOD5 proteins: mapping of sequences sufficient for targeting to mitochondria and demonstration that mitochondrial and nuclear isoforms commingle in the cytosol." Mol Cell Biol 1994;14:2298-2306.

[3] Gillman EC, Slusher LB, Martin NC, Hopper AK; Medline: 91203856 MOD5 translation initiation sites determine N6-isopentenyladenosine modification of mitochondrial
10   and cytoplasmic tRNA." Mol Cell Biol 1991;11:2382-2390.

854. (KE2)

KE2 family protein

15

The function of members of this family is unknown, although they have been suggested to contain a DNA binding leucine zipper motif [2].

Number of members: 9

20

[1] Ha H, Abe K, Artzt K; Medline: 92084131 Primary structure of the embryo-expressed gene KE2 from the mouse H-2K region." Gene 1991;107:345-346.

[2] Shang HS, Wong SM, Tan HM, Wu M; Medline: 95129859 YKE2, a yeast nuclear gene encoding a protein showing homology to mouse KE2 and containing a putative leucine-
25   zipper motif." Gene 1994;151:197-201.

855. (Lipoprotein_6)

Prokaryotic membrane lipoprotein lipid attachment site

30

Cross-reference(s) PS00013; PROKAR_LIPOPROTEIN

In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which

a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):

- Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp).

- Escherichia coli lipoprotein-28 (gene nlpA).

5 - Escherichia coli lipoprotein-34 (gene nlpB).

- Escherichia coli lipoprotein nlpC.

- Escherichia coli lipoprotein nlpD.

- Escherichia coli osmotically inducible lipoprotein B (gene osmB).

- Escherichia coli osmotically inducible lipoprotein E (gene osmE).

10 - Escherichia coli peptidoglycan-associated lipoprotein (gene pal).

- Escherichia coli rare lipoproteins A and B (genes rplA and rplB).

- Escherichia coli copper homeostasis protein cutF (or nlpE).

- Escherichia coli plasmids traT proteins.

- Escherichia coli Col plasmids lysis proteins.

15 - A number of Bacillus beta-lactamases.

- Bacillus subtilis periplasmic oligopeptide-binding protein (gene oppA).

- Borrelia burgdorferi outer surface proteins A and B (genes ospA and ospB).

- Borrelia hermsii variable major protein 21 (gene vmp21) and 7 (gene vmp7).

- Chlamydia trachomatis outer membrane protein 3 (gene omp3).

20 - Fibrobacter succinogenes endoglucanase cel-3.

- Haemophilus influenzae proteins Pal and Pcp.

- Klebsiella pullulunase (gene pulA).

- Klebsiella pullulunase secretion protein pulS.

- Mycoplasma hyorhinis protein p37.

25 - Mycoplasma hyorhinis variant surface antigens A, B, and C (genes vlpABC).

- Neisseria outer membrane protein H.8.

- Pseudomonas aeruginosa lipopeptide (gene lppL).

- Pseudomonas solanacearum endoglucanase egl.

- Rhodopseudomonas viridis reaction center cytochrome subunit (gene cytC).

30 - Rickettsia 17 Kd antigen.

- Shigella flexneri invasion plasmid proteins mxiJ and mxiM.

- Streptococcus pneumoniae oligopeptide transport protein A (gene amiA).

- Treponema pallidium 34 Kd antigen.

- Treponema pallidium membrane protein A (gene tmpA).

- Vibrio harveyi chitobiase (gene chb).

- Yersinia virulence plasmid protein yscJ.

- Halocyanin from Natrobacterium pharaonis [4], a membrane associated copper-binding protein. This is the first archaebacterial protein known to be modified in such a fashion).

From the precursor sequences of all these proteins, a consensus pattern and a set of rules to identify this type of post-translational modification were derived.

Consensus pattern: {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1)

The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least one Lys or one Arg in the first seven positions of the sequence. Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT some 100 prokaryotic proteins. Some of them are not membrane lipoproteins, but at least half of them could be.

References

[1] Hayashi S., Wu H.C., J. Bioenerg. Biomembr. 22:451-471(1990).

[2] Klein P., Somorjai R.L., Lau P.C.K., Protein Eng. 2:15-20(1988).

[3] von Heijne G., Protein Eng. 2:531-534(1989).

[4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).

856. (Lipoprotein_7)

Adhesin lipoprotein

This family consists of the p50 and variable adherence-associated antigen (Vaa) adhesins from Mycoplasma hominis. M. hominis is a mycoplasma associated with human urogenital diseases, pneumonia, and septic arthritis [1]. An adhesin is a cell surface molecule that mediates adhesion to other cells or to the surrounding surface or substrate. The Vaa antigen is a 50-kDa surface lipoprotein that has four tandem repetitive DNA sequences encoding a periodic peptide structure, and is highly immunogenic in the human host [1]. p50 is also a 50-

kDa lipoprotein, having three repeats A,B and C, that may be a tetramer of 191-kDa in its native environment [2].

Number of members: 18

[1] Zhang Q, Wise KS; Medline: 96294788   Molecular basis of size and antigenic variation of a Mycoplasma hominis adhesin encoded by divergent vaa genes.   Infect Immun 1996;64:2737-2744.

[2] Henrich B, Kitzerow A, Feldmann RC, Schaal H, Hadding U; Medline: 97047675   Repetitive elements of the Mycoplasma hominis adhesin p50 can be differentiated by monoclonal antibodies." Infect Immun 1996;64:4027-4034.

857. (MaoC_like)
MaoC like domain

The MaoC protein is found to share similarity with a wide variety of enzymes; estradiol 17 beta-dehydrogenase 4, peroxisomal hydratase-dehydrogenase-epimerase, fatty acid synthase beta subunit. All these enzymes contain other domains. This domain is also present in the NodN nodulation protein N. No specific function has been assigned to this region of any of these proteins. The maoC gene is part of a operon with maoA which is involved in the synthesis of monoamine oxidase [1].

Number of members: 46

[1] Sugino H, Sasaki M, Azakami H, Yamashita M, Murooka Y Medline: 96235221   A monoamine-regulated Klebsiella aerogenes operon containing the monoamine oxidase structural gene (maoA) and the maoC gene." J Bacteriol 1992;174:2485-2492.

858. (MSP)
Manganese-stabilizing protein / photosystem II polypeptide

This family consists of the 33 KDa photosystem II polypeptide from the oxygen evolving complex (OEC) of plants and cyanobacteria. The protein is also known as the manganese-stabilizing protein as it is associated with the manganese complex of the OEC and may provide the ligands for the complex [1].

Number of members: 17

[1] Philbrick JB, Zilinskas BA; Medline: 88334494 "Cloning, nucleotide sequence and mutational analysis of the gene encoding the Photosystem II manganese-stabilizing polypeptide of Synechocystis 6803." Mol Gen Genet 1988;212:418-425.

859. (NAC)

[1] Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV; Medline: 99342100   Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell." Genome Res 1999;9:608-628.

Number of members: 27

860. (Nop)
Putative snoRNA binding domain

This family consists of various Pre RNA processing ribonucleoproteins. The function of the aligned region is unknown however it may be a common RNA or snoRNA or Nop1p binding domain. Nop5p (Nop58p) Swiss:Q12499 from yeast is the protein component of a ribonucleoprotein protein required for pre-18s rRNA processing and is suggested to function with Nop1p in a snoRNA complex [1]. Nop56p Swiss:O00567 and Nop5p interact with Nop1p and are required for ribosome biogenesis [2]. Prp31p Swiss:p49704 is required for pre-mRNA splicing in S. cerevisiae [3].

Number of members: 23

[1] Wu P, Brockenbrough JS, Metcalfe AC, Chen S, Aris JP; Medline: 98298165 Nop5p is a small nucleolar ribonucleoprotein component required for pre- 18 S rRNA processing in yeast." J Biol Chem 1998;273:16453-16463.

[2] Gautier T, Berges T, Tollervey D, Hurt E;Medline: 8038777 Nucleolar KKE/D repeat proteins Nop56p and Nop58p interact with Nop1p and are required for ribosome biogenesis." Mol Cell Biol 1997;17:7088-7098.

[3] Weidenhammer EM, Singh M, Ruiz-Noriega M, Woolford JL Jr; Medline: 96184869 The PRP31 gene encodes a novel protein required for pre-mRNA splicing in Saccharomyces cerevisiae." Nucleic Acids Res 1996;24:1164-1170.

861. (Nramp)

Natural resistance-associated macrophage protein

The natural resistance-associated macrophage protein (NRAMP) family consists of Nramp1, Nramp2, and yeast proteins Smf1 and Smf2. The NRAMP family is a novel family of functional related proteins defined by a conserved hydrophobic core of ten transmembrane domains [5]. This family of membrane proteins are divalent cation transporters. Nramp1 is an integral membrane protein expressed exclusively in cells of the immune system and is recruited to the membrane of a phagosome upon phagocytosis [1]. By controlling divalent cation concentrations Nramp1 may regulate the interphagosomal replication of bacteria [1]. Mutations in Nramp1 may genetically predispose an individual to susceptibility to diseases including leprosy and tuberculosis conversely this might however provide protection form rheumatoid arthritis [1]. Nramp2 is a multiple divalent cation transporter for Fe2+, Mn2+ and Zn2+ amongst others it is expressed at high levels in the intestine; and is major transferrin-independent iron uptake system in mammals [1]. The yeast proteins Smf1 and Smf2 may also transport divalent cations [3].

Number of members: 36

[1] Govoni G, Gros P; Medline: 98383996 Macrophage NRAMP1 and its role in resistance to microbial infections." Inflamm Res 1998;47:277-284.

[2] Agranoff DD, Krishna S Medline: 98294035   Metal ion homeostasis and intracellular parasitism." Mol Microbiol 1998;28:403-412.

[3] Pinner E, Gruenheid S, Raymond M, Gros P; Medline: 98030569   Functional complementation of the yeast divalent cation transporter family SMF by NRAMP2, a

5   member of the mammalian natural resistance- associated macrophage protein family." J Biol Chem 1997;272:28933-28938.

[4] Cellier M, Belouchi A, Gros P; Medline: 96402487   Resistance to intracellular infections: comparative genomic analysis of Nramp." Trends Genet 1996;12:201-204.

[5] Cellier M, Prive G, Belouchi A, Kwan T, Rodrigues V, Chia W, Gros P; Medline:

10   96036029   Nramp defines a family of membrane proteins." Proc Natl Acad Sci U S A 1995;92:10089-10093.

862. (NTP_transf_2)

15   Nucleotidyltransferase domain

Members of this family belong to a large family of nucleotidyltransferases [1].

Number of members: 83

20

[1] Holm L, Sander C; Medline: 96005605   DNA polymerase beta belongs to an ancient nucleotidyltransferase superfamily." Trends Biochem Sci 1995;20:345-347.

25   863. (Paramyxo_P)

Paramyxovirus P phosphoprotein

This family consists of paramyxovirus P phosphoprotein from sendai virus and human and bovine parainfluenza viruses. The P protein is an essential part of the viral RNA polymerase

30   complex formed form the P and L proteins [1]. The exact role of the P protein in this complex in unknown but it is involved in multiple protein-protein interactions and binding the polymerase complex to the nucleocapsid or ribonucleoprotein template [1]. It also appears to be important for the proper folding of the L protein [1]. The paramyxoviruses have a negative sense ssRNA genome [1].

Number of members: 15

[1] Bowman MC, Smallwood S, Moyer SA; Medline: 99329169 Dissection of Individual Functions of the Sendai Virus Phosphoprotein in Transcription." J Virol 1999;73:6474-6483.

[2] Matsuoka Y, Curran J, Pelet T, Kolakofsky D, Ray R, Compans RW; Medline: 91237868 The P gene of human parainfluenza virus type 1 encodes P and C proteins but not a cysteine-rich V protein." J Virol 1991;65:3406-3410.

864. (Patatin)

This family consists of various patatin glycoproteins from plants. The patatin protein accounts for up to 40% of the total soluble protein in potato tubers [2]. Patatin is a storage protein but it also has the enzymatic activity of lipid acyl hydrolase, catalysing the cleavage of fatty acids from membrane lipids [2].

Number of members: 21

[1] Banfalvi Z, Kostyal Z, Barta E; Medline: 95107249 Solanum brevidens possesses a non-sucrose-inducible patatin gene." Mol Gen Genet 1994;245:517-522.

[2] Mignery GA, Pikaard CS, Park WD; Medline: 88226014 Molecular characterization of the patatin multigene family of potato." Gene 1988;62:27-44.

865. (Pentapeptide_2)
Pentapeptide repeats (8 copies)

These repeats are found in many mycobacterial proteins. These repeats are most common in the PPE family of proteins, where they are found in the MPTR subfamily of PPE proteins. The function of these repeats is unknown. The repeat can be approximately described as XNXGX, where X can be any amino acid. These repeats are similar to Pentapeptide [1], however it is not clear if these two families are structurally related.

Number of members: 362

[1] Bateman A, Murzin A, Teichmann SA; Medline: 98318059 Structure and distribution of
pentapeptide repeats in bacteria." Protein Sci 1998;7:1477-1480.

[2] Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Barrell BG; Medline: 98295987 Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence." Nature 1998;393:537-544.

866. (Peptidase_C13)

Peptidase C13 family

This family of peptidases is known as the hemoglobinase family because it contains a globin degrading enzyme from blood parasites Swiss:P42665. However relatives are found in plants and other organisms that have other functions. Members of this family are asparaginyl peptidases [1].

Number of members: 26

[1] Chen JM, Dando PM, Rawlings ND, Brown MA, Young NE, Stevens RA, Hewitt E, Watts C, Barrett AJ; Medline: 97218252 Cloning, isolation, and characterization of mammalian legumain, an asparaginyl endopeptidase." J Biol Chem 1997;272:8090-8098.

867. (Pro_dh)

Proline dehydrogenase

Number of members: 25

[1] Ling M, Allen SW, Wood JM; Medline: 95055736 Sequence analysis identifies the proline dehydrogenase and delta 1- pyrroline-5-carboxylate dehydrogenase domains of the multifunctional Escherichia coli PutA protein." J Mol Biol 1994;243:950-956.

868. (PsbP)

This family consists of the 23 kDa subunit of oxygen evolving system of photosystem II or PsbP from various plants (where it is encoded by the nuclear genome) and Cyanobacteria. The 23 KDa PsbP protein is required for PSII to be fully operational in vivo, it increases the affinity of the water oxidation site for Cl- and provides the conditions required for high affinity binding of Ca2+ [2].

Number of members: 25

[1] Rova EM, Mc Ewen B, Fredriksson PO, Styring S; Medline: 97067138 Photoactivation and photoinhibition are competing in a mutant of Chlamydomonas reinhardtii lacking the 23-kDa extrinsic subunit of photosystem II." J Biol Chem 1996;271:28918-28924.
[2] Kochhar A, Khurana JP, Tyagi AK; Medline: 97191538 Nucleotide sequence of the psbP gene encoding precursor of 23-kDa polypeptide of oxygen-evolving complex in Arabidopsis thaliana and its expression in the wild-type and a constitutively photomorphogenic mutant." DNA Res 1996;3:277-285.

869. (PUA)

The PUA domain named after PseudoUridine synthase and Archaeosine transglycosylase, was detected in archaeal and eukaryotic pseudouridine synthases, archaeal archaeosine synthases, a family of predicted ATPases that may be involved in RNA modification, a family of predicted archaeal and bacterial rRNA methylases. Additionally, the PUA domain was detected in a family of eukaryotic proteins that also contain a domain homologous to the translation initiation factor eIF1/SUI1; these proteins may comprise a novel type of translation factors. Unexpectedly, the PUA domain was detected also in bacterial and yeast glutamate kinases; this is compatible with the demonstrated role of these enzymes in the

690

regulation of the expression of other genes [1]. It is predicted that the PUA domain is an RNA binding domain.

Number of members: 48

[1] Aravind L, Koonin EV; Medline: 99193178 Novel predicted RNA-binding domains associated with the translation machinery." J Mol Evol 1999;48:291-302.

870. (RF1)

eRF1-like proteins

Members of this family are peptide chain release factors. The eukaryotic Release Factor 1 proteins (eRF1s) are involved in termination of translation. The eRF1 protein is functional for all stop codons and appears to abolish read-through of these codons. This family also includes other proteins for which the precise molecular function is unknown. Many of them are from Archaebacteria. These proteins may also be involved in translation termination but this awaits experimental verification. Number of members: 25

[1] Frolova L, Le Goff X, Rasmussen HH, Cheperegin S, Drugeon G, Kress M, Arman I, Haenni AL, Celis JE, Philippe M, et al; Medline: 95082951 A highly conserved eukaryotic protein family possessing properties of polypeptide chain release factor" [see comments] Nature 1994;372:701-703.

[2] Drugeon G, Jean-Jean O, Frolova L, Le Goff X, Philippe M, Kisselev L, Haenni AL; Medline: 97315314 Eukaryotic release factor 1 (eRF1) abolishes readthrough and competes with suppressor tRNAs at all three termination codons in messenger RNA." Nucleic Acids Res 1997;25:2254-2258.

871. (Ribosomal_L14e)Ribosomal protein L14

This family includes the eukaryotic ribosomal protein L14.

Number of members: 15

872. (Ribosomal_S27)

Ribosomal protein S27a

This family of ribosomal proteins consists mainly of the 40S ribosomal protein S27a which is synthesized as a C-terminal extension of ubiquitin (CEP). The S27a domain compromises the C-terminal half of the protein. The synthesis of ribosomal proteins as extensions of ubiquitin promotes their incorporation into nascent ribosomes by a transient metabolic stabilization and is required for efficient ribosome biogenesis [3]. The ribosomal extension protein S27a contains a basic region that is proposed to form a zinc finger; its fusion gene is proposed as a mechanism to maintain a fixed ratio between ubiquitin necessary for degrading proteins and ribosomes a source of proteins [2].

Number of members: 36

873. (Spermine_synth)

Spermine/spermidine synthase

Spermine and spermidine are polyamines. This family includes spermidine synthase that catalyses the fifth (last) step in the biosynthesis of spermidine from arginine, and spermine synthase.

Number of members: 39

[1] Mezquita J, Pau M, Mezquita C; Medline: 97449308   Characterization and expression of two chicken cDNAs encoding ubiquitin fused to ribosomal proteins of 52 and 80 amino acids." Gene 1997;195:313-319.

[2] Redman KL, Rechsteiner M; Medline: 89181932   Identification of the long ubiquitin extension as ribosomal protein S27a." Nature 1989;338:438-440.

[3] Finley D, Bartel B, Varshavsky A; Medline: 89181925   The tails of ubiquitin precursors are ribosomal proteins whose fusion to ubiquitin facilitates ribosome biogenesis." Nature 1989;338:394-401.

874. (Surp)

Surp module

[1] Denhez F, Lafyatis R; Medline: 94266805 Conservation of regulated alternative splicing
5 and identification of functional domains in vertebrate homologs to the Drosophila splicing
regulator, suppressor-of-white-apricot." J Biol Chem 1994;269:16170-16179.

This domain is also known as the SWAP domain. SWAP stands for Suppressor-of-White-
APricot. It has been suggested that these domains may be RNA binding [1].
10
Number of members: 32

875. (TFIIE)
15 TFIIE alpha subunit

The general transcription factor TFIIE has an essential role in eukaryotic transcription
initiation together with RNA polymerase II and other general factors. Human TFIIE consists
of two subunits TFIIE-alpha Swiss:P29083 and TFIIE-beta Swiss:P29084 and joins the
20 preinitiation complex after RNA polymerase II and TFIIF [1]. This family consists of the
conserved amino terminal region of eukaryotic TFIIE-alpha [2] and proteins from
archaebacteria that are presumed to be TFIIE-alpha subunits also Swiss:O29501 [3].

Number of members: 12
25

[1] Ohkuma Y, Sumimoto H, Hoffmann A, Shimasaki S, Horikoshi M, Roeder RG; Medline:
92065982 Structural motifs and potential sigma homologies in the large subunit of human
general transcription factor TFIIE." Nature 1991;354:398-401.
[2] Ohkuma Y, Hashimoto S, Roeder RG, Horikoshi M; Medline: 93087200 Identification of
30 two large subdomains in TFIIE-alpha on the basis of homology between Xenopus and human
sequences. Nucleic Acids Res 1992;20:5838-5838.
[3] Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn
M, Hickey EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC,
Fleischmann RD, Quackenbush J, Lee NH, Sutton GG, Gill S, Kirkness EF, Dougherty BA,

McKenney K, Adams MD, Loftus B, Venter JC, et al; Medline: 98049343 The complete 'genome sequence of the hyperthermophilic, sulphate- reducing archaeon Archaeoglobus fulgidus." Nature 1997;390:364-370.

5

876. (Transglut_core)

Cross-reference(s) PS00547; TRANSGLUTAMINASES

10   Transglutaminases (EC 2.3.2.13) (TGase) [1,2] are calcium-dependent enzymes that catalyze the cross-linking of proteins by promoting the formation of isopeptide bonds between the gamma-carboxyl group of a glutamine in one polypeptide chain and the epsilon-amino group of a lysine in a second polypeptide chain. TGases also catalyze the conjugation of polyamines to proteins. The best known transglutaminase is blood coagulation factor XIII, a plasma

15   tetrameric protein composed of two catalytic A subunits and two non-catalytic B subunits. Factor XIII is responsible for cross-linking fibrin chains, thus stabilizing the fibrin clot. Other forms of transglutaminases are widely distributed in various organs, tissues and body fluids. Sequence data is available for the following forms of TGase:

- Transglutaminase K (Tgase K), a membrane-bound enzyme found in mammalian epidermis

20   and important for the formation of the cornified cell envelope (gene TGM1).

- Tissue transglutaminase (TGase C), a monomeric ubiquitous enzyme located in the cytoplasm (gene TGM2).

- Transglutaminase 3, responsible for the later stages of cell envelope formation in the epidermis and the hair follicle (gene TGM3).

25   - Transglutaminase 4 (gene TGM4).

A conserved cysteine is known to be involved in the catalytic mechanism of TGases. The erythrocyte membrane band 4.2 protein, which probably plays an important role in regulating the shape of erythrocytes and their mechanical properties, is evolutionary related to TGases.

30   However the active site cysteine is substituted by an alanine and the 4.2 protein does not show TGase activity.

Consensus pattern:[GT]-Q-[CA]-W-V-x-[SA]-[GA]-[IVT]-x(2)-T-x-[LMSC]-R-[CSA]-[LV]-G [The first C is the active site residue]   Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROTNONE.

5   [1] Ichinose A., Bottenus R.E., Davie E.W. J. Biol. Chem. 265:13411-13414(1990).

[2] Greenberg C.S., Birckbichler P.J., Rice R.H. FASEB J. 5:3071-3077(1991).

877. (TruB_N)

10   TruB family pseudouridylate synthase (N terminal domain)

Members of this family are involved in modifying bases in RNA molecules. They carry out the conversion of uracil bases to pseudouridine. This family includes TruB, a pseudouridylate synthase that specifically converts uracil 55 to pseudouridine in most tRNAs. This family

15   also includes Cbf5p that modifies rRNA [2].

Number of members: 33

[1] Nurse K, Wrzesinski J, Bakin A, Lane BG, Ofengand J; Medline: 96079944   Purification,

20   cloning, and properties of the tRNA psi 55 synthase from Escherichia coli." RNA 1995;1:102-112.

[2] Lafontaine DLJ, Bousquet-Antonelli C, Henry Y, Caizergues-Ferrer M, Tollervey D; Medline: 98139521   The box H + ACA snoRNAs carry Cbf5p, the putative rRNA pseudouridine synthase." Genes Dev 1998;12:527-537.

25

878. (UDPGP)

UTP--glucose-1-phosphate uridylyltransferase

30   This family consists of UTP--glucose-1-phosphate uridylyltransferases, EC:2.7.7.9. Also known as UDP-glucose pyrophosphorylase (UDPGP) and Glucose-1-phosphate uridylyltransferase. UTP--glucose-1-phosphate uridylyltransferase catalyses the interconversion of MgUTP + glucose-1-phosphate and UDP-glucose + MgPPi [1]. UDP-glucose is an important intermediate in mammalian carbohydrate interconversion involved in

various metabolic roles depending on tissue type [1]. In Dictyostelium (slime mold) mutants in this enzyme abort the development cycle [2]. Also within the family is UDP-N-acetylglucosamine Swiss:Q16222 or AGX1 [3] and two hypothetical proteins from Borrelia burgdorferi the lyme disease spirochaete Swiss:O51893 and Swiss:O51036.

Number of members: 18

[1] Duggleby RG, Chao YC, Huang JG, Peng HL, Chang HY; Medline: 96202932 Sequence differences between human muscle and liver cDNAs for UDPglucose pyrophosphorylase and kinetic properties of the recombinant enzymes expressed in Escherichia coli." Eur J Biochem 1996;235:173-179.

[2] Ragheb JA, Dottin RP; Medline: 87231075 Structure and sequence of a UDP glucose pyrophosphorylase gene of Dictyostelium discoideum." Nucleic Acids Res 1987;15:3891-3906.

[3] Mio T, Yabe T, Arisawa M, Yamada-Okabe H; Medline: 98269105 The eukaryotic UDP-N-acetylglucosamine pyrophosphorylases. Gene cloning, protein expression, and catalytic mechanism. J Biol Chem 1998;273:14392-14397.

879. (UPF004)

Uncharacterized protein family UPF0044 signature
Cross-reference(s) PS01301; UPF0044

The following uncharacterized proteins have been shown [1] to be highlysimilar:
- Bacillus subtilis hypothetical protein yqeI.
- Escherichia coli hypothetical protein yhbY and HI1333, the corresponding Haemophilus influenzae protein.
- Methanococcus jannaschii hypothetical protein MJ0652.
These are small proteins of 10 to 15 Kd. They can be picked up in the database by the following pattern. This pattern is located in the N-terminal part of these proteins.

Consensus pattern: L-[ST]-x(3)-K-x(3)-[KR]-[SGA]-x-[GA]-H-x-L-x-P-[LIV]-x(2)- [LIV]-[GA]-x(2)-G Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROTNONE.

5

880. (zf-A20)

A20-like zinc finger

A20- (an inhibitor of cell death)-like zinc fingers. The zinc

finger mediates self-association in A20. These fingers also

10   mediate IL-1-induced NF-kappa B activation.

Number of members: 22

[1] Heyninck K, Beyaert R; Medline: 99126071   The cytokine-inducible zinc finger protein

15   A20 inhibits IL-1-induced NF- kappaB activation at the level of TRAF6. FEBS Lett

1999;442:147-150.

[2] De Valck D, Heyninck K, Van Criekinge W, Contreras R,Beyaert R, Fiers W; Medline:

96390831   A20, an inhibitor of cell death, self-associates by its

zinc finger domain." FEBS Lett 1996;384:61-64.

20   [3] Song HY, Rothe M, Goeddel DV; Medline: 96270609   The tumor necrosis factor-

inducible zinc finger protein A20 interacts with TRAF1/TRAF2 and inhibits NF-kappaB

activation. Proc Natl Acad Sci U S A 1996;93:6721-6725.

[4] Opipari AW Jr, Boguski MS, Dixit VM; Medline: 90368626   The A20 cDNA induced by

tumor necrosis factor alpha encodes a novel type of zinc finger protein." J Biol Chem

25   1990;265:14705-14708.

881. (zf-PARP)

Poly(ADP-ribose) polymerase zinc finger domain

30

Cross-reference(s) PS00347; PARP_ZN_FINGER_1 PS50064; PARP_ZN_FINGER_2

Poly(ADP-ribose) polymerase (EC 2.4.2.30) (PARP) [1,2] is a eukaryotic enzyme that catalyzes the covalent attachment of ADP-ribose units from NAD(+) to various nuclear

acceptor proteins. This post-translational modification of nuclear proteins is dependent on DNA. It appears to be involved in the regulation of various important cellular processes such as differentiation, proliferation and tumor transformation as well as in the regulation of the molecular events involved in the recovery of the cell from DNA damage.

5   Structurally, PARP, about 1000 amino-acids residues long, consists of three distinct domains: an N-terminal zinc-dependent DNA-binding domain, a central automodification domain and a C-terminal NAD-binding domain. The DNA-binding region contains a pair of zinc finger domains which have been shown to bind DNA in a zinc-dependent manner. The zinc finger domains of PARP seem to bind specifically to single-stranded DNA. DNA ligase

10   III [3] contains, in its N-terminal section, a single copy of a zinc finger highly similar to those of PARP.

Consensus pattern: C-[KR]-x-C-x(3)-I-x-K-x(3)-[RG]-x(16,18)-W-[FYH]-H-x(2)-C [The three C's and the H are zinc ligands] Sequences known to belong to this class detected by the

15   patternALL. Other sequence(s) detected in SWISS-PROTNONE. Sequences known to belong to this class detected by the profile ALL. Other sequence(s) detected in SWISS-PROTNONE.

Note: This documentation entry is linked to both signature patterns and a profile. As the

20   profile is much more sensitive than the patterns, you should use it if you have access to the necessary software tools to do so.

[1] Althaus F.R., Richter C.R. Mol. Biol. Biochem. Biophys. 37:1-126(1987).

[2] de Murcia G., Menissier de Murcia J. Trends Biochem. Sci. 19:172-176(1994).

25   [3] Wei Y.-F., Robins P., Carter K., Caldecott K., Pappin D.J.C., Yu G.-L., Wang R.-P., Shell B.K., Nash R.A., Schar P., Barnes D.E., Haseltine W.A., Lindahl T. Mol. Cell. Biol. 15:3206-3216(1995).

882. Adenylylsulfate kinase (APS_kinase)

30   Enzyme that catalyses the phosphorylation of adenylylsulfate to 3'-phosphoadenylylsulfate. This domain contains an ATP binding P-loop motif. Number of members: 34

698

[1] MacRae IJ, Rose AB, Segel IH; Medline: 99003196 Adenosine 5'-phosphosulfate kinase from Penicillium chrysogenum. site- directed mutagenesis at putative phosphoryl-accepting and ATP P-loop residues. J Biol Chem 1998;273:28583-28589.

5

883. DNA polymerase family B signature DNA_POLYMERASE_B (DNA_pol_B)

Replicative DNA polymerases (EC 2.7.7.7) are the key enzymes catalyzing the accurate replication of DNA. They require either a small RNA molecule or a protein as a primer for the de novo synthesis of a DNA chain. On the basis of sequence similarity, a

10 number of DNA polymerases have been grouped [1 to 7] under the designation of DNA polymerase family B. These are:

- Higher eukaryotes polymerases alpha.

- Higher eukaryotes polymerases delta.

- Yeast polymerase I/alpha (gene POL1), polymerase II/epsilon (gene POL2), polymerase

15 III/delta (gene POL3) and polymerase REV3.

- Escherichia coli polymerase II (gene dinA or polB).

- Archaebacterial polymerases.

- Polymerases of viruses from the herpesviridae family.

- Polymerases from Adenoviruses.

20 - Polymerases from Baculoviruses.

- Polymerases from Chlorella viruses.

- Polymerases from Poxviruses.

- Bacteriophage T4 polymerase.

- Podoviridae bacteriophages Phi-29, M2 and PZA polymerase.

25 - Tectiviridae bacteriophage PRD1 polymerase.

- Polymerases encoded on mitochondrial linear DNA plasmids in various fungi and plants (Kluyveromyces lactis pGKL1 and pGKL2, Agaricus bitorquis pEM, Ascobolus immersus pAI2, Claviceps purpurea pCLK1, Neurospora Kalilo and Maranhar, maize S-1, etc).

30 Six regions of similarity (numbered from I to VI) are found in all or a subset of the above polymerases. The most conserved region (I) includes a conserved tetrapeptide with two aspartate residues. Its function is not yet known. However, it has been suggested [3] that it may be involved in binding a magnesium ion. This conserved region was selected as a signature for this family of DNA polymerases.

Consensus pattern [YA]-[GLIVMSTAC]-D-T-D-[SG]-[LIVMFTC]-x-[LIVMSTAC]

Sequences known to belong to this class detected by the patternALL, except for yeast polymerase II/epsilon, Agaricus bitorquis pEM and Sulfolobus solfataricus polymerase II.

5

[1] Jung G., Leavitt M.C., Hsieh J.-C., Ito J. Proc. Natl. Acad. Sci. U.S.A. 84:8287-8291(1987).

[2] Bernad A., Zaballos A., Salas M., Blanco L. EMBO J. 6:4219-4225(1987).

[3] Argos P. Nucleic Acids Res. 16:9909-9916(1988).

10   [4] Wang T.S.-F., Wong S.W., Korn D. FASEB J. 3:14-21(1989).

[5] Delarue M., Poch O., Todro N., Moras D., Argos P. Protein Eng. 3:461-467(1990).

[6] Ito J., Braithwaite D.K. Nucleic Acids Res. 19:4045-4057(1991).

[7] Braithwaite D.K., Ito J. Nucleic Acids Res. 21:787-802(1993).

15

884. DNA polymerase family X signature - DNA_POLYMERASE_X (DNA_polymeraseX)

DNA polymerases (EC 2.7.7.7) can be classified, on the basis of sequence similarity [1], into at least four different groups: A, B, C and X. DNA polymerases that belong to family X are
20   listed below [2]:

- Vertebrate polymerase beta, involved in DNA repair.

- Yeast polymerase IV (POL4) [3], an enzyme with similar characteristics to that of the mammalian polymerase beta.

- Terminal deoxynucleotidyltransferase (TdT) (EC 2.7.7.31). TdT catalyzes the elongation of
25   polydeoxynucleotide chains by terminal addition. One of the functions of this enzyme is the addition of nucleotides at the junction of rearranged Ig heavy chain and T cell receptor gene segments during the maturation of B and T cells.

- African Swine Fever virus protein O174L [4].

- Fission yeast hypothetical protein SpAC2F7.06c.

30

These enzymes are small (about 40 Kd) compared with other polymerases and their reaction mechanism operates via a distributive mode, i.e. they dissociate from the template-primer after addition of each nucleotide.

As a signature pattern for this family of DNA polymerases, a highly conserved region that contains a conserved arginine and two conserved aspartic acid residues were selected. The latter together with the arginine have been shown [5] to be involved in primer binding in polymerase beta.

Consensus pattern G-[SG]-[LFY]-x-R-[GE]-x(3)-[SGCL]-x-D-[LIVM]-D- [LIVMFY](3)-x(2)-[SAP] Sequences known to belong to this class detected by the patternALL.

[1] Ito J., Braithwaite D.K. Nucleic Acids Res. 19:4045-4057(1991).

[2] Matsukage A., Nishikawa K., Ooi T., Seto Y., Yamaguchi M. J. Biol. Chem. 262:8960-8962(1987).

[3] Prasad R., Widen S.G., Singhal R.K., Watkins J., Prakash L., Wilson S.H. Nucleic Acids Res. 21:5301-5307(1993).

[4] Yanez R.J., Rodriguez J.M., Nogal M.L., Yuste L., Enriquez C., Rodriguez J.F., Vinuela E. Virology 208:249-278(1995).

[5] Date T., Yamamoto S., Tanihara K., Nishimoto Y., Matsukage A. Biochemistry 30:5286-5292(1991).

885. DUF14 - Domain of unknown function

This domain is found in glutamate synthase, tungsten formylmethanofuran dehydrogenase subunit c (FwdC) and molybdenum formylmethanofuran dehydrogenase subunit c (FmdC). It has no known function. Number of members: 52

[1] Hochheimer A, Hedderich R, Thauer RK; Medline: 99035764. The formylmethanofuran dehydrogenase isoenzymes in Methanobacterium wolfei and Methanobacterium thermoautotrophicum: induction of the molybdenum isoenzyme by molybdate and constitutive synthesis of the tungsten isoenzyme." Arch Microbiol 1998;170:389-393.

886. DUF18-Domain of unknown function

This domain of unknown function is found in several C. elegans proteins. The domain is 120 amino acids long and rich in cysteine residues. There are 16 conserved cysteine positions in the domain. Number of members: 34

887. DUF27-Domain of unknown function

This domain is found in a number of otherwise unrelated proteins. This domain is found at the C-terminus of the macro-H2A histone protein Swiss:Q02874. This domain is found in the non-structural proteins of several types of ssRNA viruses such as NSP2 from alphaviruses Swiss:P03317. This domain is also found on its own in a family of proteins from bacteria Swiss:P75918, archaebacteria Swiss:O59182 and eukaryotes Swiss:Q17432, suggesting that it is involved in an important and ubiquitous cellular process. Number of members: 66

888. DUF37-Domain of unknown function

This domain is found in short (70 amino acid) hypothetical proteins from various bacteria. The domain contains three conserved cysteine residues. Swiss:Q44066 from Aeromonas hydrophila has been found to have hemolytic activity (unpublished). Number of members: 19

889. EGF-like domain signatures. (EGF-like)

A sequence of about thirty to forty amino-acid residues long found in the sequence of epidermal growth factor (EGF) has been shown [1 to 6] to be present, in a more or less conserved form, in a large number of other, mostly animal proteins. The proteins currently known to contain one or more copies of an EGF-like pattern are listed below.

- Adipocyte differentiation inhibitor (gene PREF-1) from mouse (6 copies).
- Agrin, a basal lamina protein that causes the aggregation of acetylcholine receptors on cultured muscle fibers (4 copies).
- Amphiregulin, a growth factor (1 copy).
- Betacellulin, a growth factor (1 copy).
- Blastula proteins BP10 and Span from sea urchin which are thought to be involved in pattern formation (1 copy).
- BM86, a glycoprotein antigen of cattle tick (7 copies).
- Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation and which expresses metalloendopeptidase activity (1-2 copies). Homologous proteins are found in sea urchin - suBMP (1 copy) - and in Drosophila - the dorsal-ventral patterning protein tolloid (2 copies).
- Caenorhabditis elegans developmental proteins lin-12 (13 copies) and glp-1 (10 copies).
- Caenorhabditis elegans APX-1 protein, a patterning protein (4.5 copies).
- Calcium-dependent serine proteinase (CASP) which degrades the extracellular matrix proteins type I and IV collagen and fibronectin (1 copy).

- Cartilage matrix protein CMP (1 copy).

- Cartilage oligomeric matrix protein COMP (4 copies).

- Cell surface antigen 114/A10 (3 copies).

- Cell surface glycoprotein complex transmembrane subunit ASGP-2 from rat (2 copies).

5   - Coagulation associated proteins C, Z (2 copies) and S (4 copies).

- Coagulation factors VII, IX, X and XII (2 copies).

- Complement C1r components (1 copy).

- Complement C1s components (1 copy).

- Complement-activating component of Ra-reactive factor (RARF) (1 copy).

10   - Complement components C6, C7, C8 alpha and beta chains, and C9 (1 copy).

- Crumbs, an epithelial development protein from Drosophila (29 copies).

- Epidermal growth factor precursor (7-9 copies).

- Exogastrula-inducing peptides A, C, D and X from sea urchin (1 copy).

- Fat protein, a Drosophila cadherin-related tumor suppressor (5 copies).

15   - Fetal antigen 1, a probable neuroendocrine differentiation protein, which is derived from

the delta-like protein (DLK) (6 copies).

- Fibrillin 1 (47 copies) and fibrillin 2 (14 copies).

- Fibropellins IA (21 copies), IB (13 copies), IC (8 copies), II (4 copies) and III (8 copies)

from the apical lamina - a component of the extracellular matrix - of sea urchin.

20   - Fibulin-1 and -2, two extracellular matrix proteins (9-11 copies).

- Giant-lens protein (protein Argos), which regulates cell determination and axon guidance in

the Drosophila eye (1 copy).

- Growth factor-related proteins from various poxviruses (1 copy).

- Gurken protein, a Drosophila developmental protein (1 copy).

25   - Heparin-binding EGF-like growth factor (HB-EGF), transforming growth factor alpha

(TGF-alpha), growth factors Lin-3 and Spitz (1 copy); the precursors are membrane proteins,

the mature form is located extracellular.

- Hepatocyte growth factor (HGF) activator (EC 3.4.21.-) (2 copies).

- LDL and VLDL receptors, which bind and transport low-density lipoproteins and very low-

30   density lipoproteins (3 copies).

- LDL receptor-related protein (LRP), which may act as a receptor for endocytosis of

extracellular ligands (22 copies).

- Leucocyte antigen CD97 (3 copies), cell surface glycoprotein EMR1 (6 copies) and cell

surface glycoprotein F4/80 (7 copies).

703

- Limulus clotting factor C, which is involved in hemostasis and host defense mechanisms in japanese horseshoe crab (1 copy).

- Meprin A alpha subunit, a mammalian membrane-bound endopeptidase (1 copy).

- Milk fat globule-EGF factor 8 (MFG-E8) from mouse (2 copies).

5 - Neuregulin GGF-I and GGF-II, two human glial growth factors (1 copy).

- Neurexins from mammals (3 copies).

- Neurogenic proteins Notch, Xotch and the human homolog Tan-1 (36 copies), Delta (9 copies) and the similar differentiation proteins Lag-2 from Caenorhabditis elegans (2 copies), Serrate (14 copies) and Slit (7 copies) from Drosophila.

10 - Nidogen (also called entactin), a basement membrane protein from chordates (2-6 copies).

- Ookinete surface proteins (24 Kd, 25 Kd, 28 Kd) from Plasmodium (4 copies).

- Pancreatic secretory granule membrane major glycoprotein GP2 (1 copy).

- Perforin, which lyses non-specifically a variety of target cells (1 copy).

- Proteoglycans aggrecan (1 copy), versican (2 copies), perlecan (at least 2 copies), brevican

15 (1 copy) and chondroitin sulfate proteoglycan (gene PG-M) (2 copies).

- Prostaglandin G/H synthase 1 and 2 (EC 1.14.99.1) (1 copy), which is found in the endoplasmatic reticulum.

- S1-5, a human extracellular protein whose ultimate activity is probably modulated by the environment (5 copies).

20 - Schwannoma-derived growth factor (SDGF), an autocrine growth factor as well as a mitogen for different target cells (1 copy).

- Selectins. Cell adhesion proteins such as ELAM-1 (E-selectin), GMP-140 (P-selectin), or the lymph-node homing receptor (L-selectin) (1 copy).

- Serine/threonine-protein kinase homolog (gene Pro25) from Arabidopsis thaliana, which

25 may be involved in assembly or regulation of light-harvesting chlorophyll A/B protein (2 copies).

- Sperm-egg fusion proteins PH-30 alpha and beta from guinea pig (1 copy).

- Stromal cell derived protein-1 (SCP-1) from mouse (6 copies).

- TDGF-1, human teratocarcinoma-derived growth factor 1 (1 copy).

30 - Tenascin (or neuronectin), an extracellular matrix protein from mammals (14.5 copies), chicken (TEN-A) (13.5 copies) and the related proteins human tenascin-X (18 copies) and tenascin-like proteins TEN-A and TEN-M from Drosophila (8 copies).

- Thrombomodulin (fetomodulin), which together with thrombin activates protein C (6 copies).

- Thrombospondin 1, 2 (3 copies), 3 and 4 (4 copies), adhesive glycoproteins that mediate cell-to-cell and cell-to-matrix interactions.

- Thyroid peroxidase 1 and 2 (EC 1.11.1.8) from human (1 copy).

- Transforming growth factor beta-1 binding protein (TGF-B1-BP) (16 or 18 copies).

5   - Tyrosine-protein kinase receptors Tek and Tie (EC 2.7.1.112) (3 copies).

- Urokinase-type plasminogen activator (EC 3.4.21.73) (UPA) and tissue plasminogen activator (EC 3.4.21.68) (TPA) (1 copy).

- Uromodulin (Tamm-horsfall urinary glycoprotein) (THP) (3 copies).

- Vitamin K-dependent anticoagulants protein C (2 copies) and protein S (4 copies) and the

10   similar protein Z, a single-chain plasma glycoprotein of unknown function (2 copies).

- 63 Kd sperm flagellar membrane protein from sea urchin (3 copies).

- 93 Kd protein (gene nel) from chicken (5 copies).

- Hypothetical 337.6 Kd protein T20G5.3 from Caenorhabditis elegans (44 copies).

15   The functional significance of EGF domains in what appear to be unrelated proteins is not yet clear. However, a common feature is that these repeats are found in the extracellular domain of membrane-bound proteins or in proteins known to be secreted (exception: prostaglandin G/H synthase). The EGF domain includes six cysteine residues which have been shown (in EGF) to be involved in disulfide bonds. The main structure is a two-stranded beta-sheet

20   followed by a loop to a C-terminal short two-stranded sheet. Subdomains between the conserved cysteines strongly vary in length as shown in the following schematic representation of the EGF-like domain:

```
        +------------------+        +-----------------------+        |           |    |
      | x(4)-C-x(0,48)-C-x(3,12)-C-x(1,70)-C-x(1,6)-C-x(2)-G-a-x(0,21)-G-x(2)-C-x     |
25    |        ********************************************
         +------------------+
```

'C': conserved cysteine involved in a disulfide bond.

'G': often conserved glycine

30   'a': often conserved aromatic amino acid

'*': position of both patterns.

'x': any residue

The region between the 5th and 6th cysteine contains two conserved glycines of which at least one is present in most EGF-like domains. Two patterns were created for this domain, each including one of these C-terminal conserved glycine residues.

5   Consensus pattern: C-x-C-x(5)-G-x(2)-C [The 3 C's are involved in disulfide bonds] Sequences known to belong to this class detected by the pattern A majority, but not those that have very long or very short regions between the last 3 conserved cysteines of their EGF-like domain(s). Other sequence(s) detected in SWISS-PROT87 proteins, of which 27 can be considered as possible candidates.

10

Consensus pattern: C-x-C-x(2)-[GP]-[FYW]-x(4,8)-C [The three C's are involved in disulfide bonds]Sequences known to belong to this class detected by the patternA majority, but not those that have very long or very short regions between the last 3 conserved cysteines of their EGF-like domain(s). Other sequence(s) detected in SWISS-PROT83 proteins, of which 49

15   can be considered as possible candidates. Note The beta chain of the integrin family of proteins contains 2 cysteine- rich repeats which were said to be dissimilar with the EGF pattern [7].

Note Laminin EGF-like repeats (see <PDOC00961>) are longer than the average EGF

20   module and contain a further disulfide bond C-terminal of the EGF-like region. Perlecan and agrin contain both EGF-like domains and laminin-type EGF-like domains. Note the pattern do not detect all of the repeats of proteins with multiple EGF-like repeats. Note see <PDOC00913> for an entry describing specifically the subset of EGF- like domains that bind calcium.

25

[1] Davis C.G. New Biol. 2:410-419(1990).

[2] Blomquist M.C., Hunt L.T., Barker W.C. Proc. Natl. Acad. Sci. U.S.A. 81:7363-7367(1984).

[3] Barker W.C., Johnson G.C., Hunt L.T., George D.G. Protein Nucl. Acid Enz. 29:54-

30   68(1986).

[4] Doolittle R.F., Feng D.F., Johnson M.S. Nature 307:558-560(1984).

[5] Appella E., Weber I.T., Blasi F. FEBS Lett. 231:1-4(1988).

[6] Campbell I.D., Bork P. Curr. Opin. Struct. Biol. 3:385-392(1993).

[7] Tamkun J.W., DeSimone D.W., Fonda D., Patel R.S., Buck C., Horwitz A.F., Hynes R.O. Cell 46:271-282(1986).

5   890. Ham1 family (Ham1p_like)

This family consists of the HAM1 protein Swiss:P47119 and hypothetical archaeal bacterial and C. elegans proteins. HAM1 controls 6-N-hydroxylaminopurine (HAP) sensitivity and mutagenesis in S. cerevisiae Swiss:P47119 [1]. The HAM1 protein protects the cell from HAP, either on the level of deoxynucleoside triphosphate or the DNA level by a yet

10   unidentified set of reactions [1]. Number of members: 19

[1] Noskov VN, Staak K, Shcherbakova PV, Kozmin SG, Negishi K, Ono BC, Hayatsu H, Pavlov YI; Medline: 96381244 HAM1, the gene controlling 6-N-hydroxylaminopurine sensitivity and mutagenesis in the yeast Saccharomyces cerevisiae." Yeast 1996;12:17-29.

15

891. (HCO3_cotransp)

Anion exchange is a cellular transport function which contributes to the regulation of cell pH and volume. Anion exchangers are a family of functionally related proteins that contributes to

20   these properties by maintaining the intracellular level of the two principal anions: chloride and HCO3-. The best characterized anion exchanger is the band 3 protein [1], which is an erythrocyte anion exchange membrane glycoprotein. Band 3 is a protein of about 900 amino acids which consists of a cytoplasmic N-terminal domain of about 400 residues and an hydrophobic C-terminal section of about 500 residues that contains at least ten

25   transmembrane regions. The cytoplasmic domain provides binding sites for cytoskeletal proteins, while the integral membrane domain is responsible for anion transport. Band 3 protein is specific to erythroid cells, at least two other proteins [2] structurally and functionally related to band 3, are found in nonerythroid tissues:

- AE2 (or B3 related protein; B3RP), a protein of 1200 residues, which seems to be present

30   in a variety of cell types including lymphoid, kidney, and choroid plexus.

- AE3, a protein of 1200 residues, which is specific to neurons.

Structurally AE2 and AE3 are very similar to band 3, the main difference being an extension of some 300 residues of the N-terminal domain in AE2 and AE3.

Two signature patterns were developed for these proteins. The first pattern is based on a conserved stretch of sequence that contains four clustered positive charged residues and which is located at the C-terminal extremity of the cytoplasmic domain, just before the first transmembrane segment from the integral domain. The second pattern is based on the

5 perfectly conserved sequence of the fifth transmembrane segment; this segment contains a lysine, which is the covalent binding site for the isothiocyanate group of DIDS, an inhibitor of anion exchange.

Consensus pattern F-G-G-[LIVM](2)-[KR]-D-[LIVM]-[RK]-R-R-Y Sequences known to

10 belong to this class detected by the pattern ALL.

Consensus pattern [FI]-L-I-S-L-I-F-I-Y-E-T-F-x-K-L Sequences known to belong to this class detected by the pattern ALL.

15 [1] Jay D., Cantley L. Annu. Rev. Biochem. 55:511-538(1986).
[2] Reithmeier R.A.F. Curr. Opin. Struct. Biol. 3:515-523(1993).

892. ATP phosphoribosyltransferase signature (HisG)

20 ATP phosphoribosyltransferase (EC 2.4.2.17) is the enzyme that catalyzes the first step in the biosynthesis of histidine in bacteria, fungi and plants. It is a protein of about 23 to 32 Kd. As a signature pattern a region located in the C-terminal part of this enzyme was selected.

Consensus pattern E-x(5)-G-x-[SAG]-x(2)-[IV]-x-D-[LIV]-x(2)-[ST]-G-x-T-[LM]

25 Sequences known to belong to this class detected by the pattern ALL.

893. HNH endonuclease (HNH)

Number of members: 56

30 [1] Shub DA, Goodrich-Blair H, Eddy SR; Medline: 95117127 Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns."
Trends Biochem Sci 1994;19:402-404.

[2] Dalgaard JZ, Klar AJ, Moser MJ, Holley WR, Chatterjee A, Mian IS; Medline: 98026854
Statistical modeling and analysis of the LAGLIDADG family of site- specific endonucleases

and identification of an intein that encodes a site-specific endonuclease of the HNH family." Nucleic Acids Res 1997;25:4626-4638.

[3] Gorbalenya AE; Medline: 95004046 Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family." Protein Sci 1994;3:1117-1120.

5

894. NEUROHYPOPHYS_HORM (hormone5)

Oxytocin (or ocytocin) and vasopressin [1] are small (nine amino acid residues), structurally and functionally related neurohypophysial peptide hormones. Oxytocin causes contraction of the smooth muscle of the uterus and of the mammary gland while vasopressin has a direct

10 antidiuretic action on the kidney and also causes vasoconstriction of the peripheral vessels. Like the majority of active peptides, both hormones are synthesized as larger protein precursors that are enzymatically converted to their mature forms. Peptides belonging to this family are also found in birds, fish, reptiles and amphibians (mesotocin, isotocin, valitocin, glumitocin, aspargtocin, vasotocin, seritocin, asvatocin, phasvatocin), in worms (annetocin),

15 octopi (cephalotocin), locust (locupressin or neuropeptide F1/F2) and in molluscs (conopressins G and S) [2]. The pattern developed to detect this category of peptides spans their entire sequence and includes four invariant amino acid residues.

Consensus pattern C-[LIFY](2)-x-N-[CS]-P-x-G [The two C's are linked by a disulfide

20 bond]. Sequences known to belong to this class detected by the pattern ALL.

[1] Acher R., Chauvet J. Biochimie 70:1197-1207(1988).

[2] Chauvet J., Michel G., Ouedraogo Y., Chou J., Chait B.T., Acher R. Int. J. Pept. Protein Res. 45:482-487(1995).

25

895. 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (HPPK)

All organisms require reduced folate cofactors for the synthesis of a variety of metabolites. Most microorganisms must synthesize folate de novo because they lack the active transport

30 system of higher vertebrate cells which allows these organisms to use dietary folates. Enzymes involved in folate biosynthesis are therefore targets for a variety of antimicrobial agents such as trimethoprim or sulfonamides. 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (EC 2.7.6.3) (HPPK) catalyzes the attachment of pyrophosphate to 6-hydroxymethyl-7,8-dihydropterin to form 6-hydroxymethyl-7,8-dihydropteridine

pyrophosphate. This is the first step in a three-step pathway leading to 7,8-dihydrofolate. Bacterial HPPK (gene folK or sulD) [1] is a protein of 160 to 270 amino acids. In the lower eukaryote Pneumocystis carinii, HPPK is the central domain of a multifunctional folate synthesis enzyme (gene fas) [2]. As a signature for HPPK, a conserved region located in the

5   central section of these enzymes was selected.

Consensus pattern [KRHD]-x-[GA]-[PSAE]-R-x(2)-D-[LIV]-D-[LIVM](2) Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROTNONE.

10

[1] Talarico T.L., Ray P.H., Dev I.K., Merrill B.M., Dallas W.S.   J. Bacteriol. 174:5971-5977(1992).

[2] Volpes F., Dyer M., Scaife J.G., Darby G., Stammers D.K., Delves C.J.   Gene 112:213-218(1992).

15

896. Metalloenzyme superfamily (Metalloenzyme)

This family includes phosphopentomutase Swiss:P07651 and 2,3-bisphosphoglycerate-independent phosphoglycerate mutase, Swiss:P37689. This family is also related to

20   alk_phosphatase [1]. The alignment contains the most conserved residues that are probably involved in metal binding and catalysis. Number of members: 34

[1] Galperin MY, Bairoch A, Koonin EV; Medline: 99180418   A superfamily of metalloenzymes unifies phosphopentomutase and cofactor- independent phosphoglycerate

25   mutase with alkaline phosphatases and sulfatases." Protein Sci 1998;7:1829-1835.

897. Penicillin amidase (Penicil_amidase)

Penicillin amidase or penicillin acylase EC:3.5.1.11 catalyses the hydrolysis of

30   benzylpenicillin to phenylacetic acid and 6-aminopenicillanic acid (6-APA) a key intermediate in the the synthesis of penicillins [1]. Also in the family is cephalosporin acylase Swiss:P07662 and Swiss:P29958 aculeacin A acylase which are involved in the synthesis of related peptide antibiotics. Number of members: 13

[1] Verhaert RM, Riemens AM, van der Laan JM, van Duin J, Quax WJ; Medline: 97438505
Molecular cloning and analysis of the gene encoding the thermostable penicillin G acylase
from Alcaligenes faecalis. Appl Environ Microbiol 1997;63:3412-3418.

[2] Duggleby HJ, Tolley SP, Hill CP, Dodson EJ, Dodson G, Moody PC; Medline: 95115804
Penicillin acylase has a single-amino-acid catalytic centre." Nature 1995;373:264-268.

898. Phosphoribosyl-AMP cyclohydrolase (PRA-CH)

This enzyme catalyses the third step in the histidine biosynthetic pathway. It requires Zn ions
for activity. Number of members: 13

[1] D'Ordine RL, Klem TJ, Davisson VJ; Medline: 99129952 N1-(5'-
phosphoribosyl)adenosine-5'-monophosphate cyclohydrolase: purification and
characterization of a unique metalloenzyme. Biochemistry 1999;38:1537-1546.

899. Phosphoribosyl-ATP pyrophosphohydrolase (PRA-PH)

This enzyme catalyses the second step in the histidine biosynthetic pathway. Number of
members: 32

[1] Keesey JK Jr, Bigelis R, Fink GR; Medline: 79216449 The product of the his4 gene
cluster in Saccharomyces cerevisiae. A trifunctional polypeptide." J Biol Chem 1979 Aug
10;254:7427-7433.

[2] Bruni CB, Carlomagno MS, Formisano S, Paolella G; Medline: 86310274 Primary and
secondary structural homologies between the HIS4 gene product of Saccharomyces
cerevisiae and the hisIE and hisD gene products of Escherichia coli and Salmonella
typhimurium." Mol Gen Genet 1986;203:389-396.

900. Prokaryotic membrane lipoprotein lipid attachment site (PstS)

In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which
is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase
recognizes a conserved sequence and cuts upstream of a cysteine residue to which a

711

glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):

- Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp).

- Escherichia coli lipoprotein-28 (gene nlpA).

5 - Escherichia coli lipoprotein-34 (gene nlpB).

- Escherichia coli lipoprotein nlpC.

- Escherichia coli lipoprotein nlpD.

- Escherichia coli osmotically inducible lipoprotein B (gene osmB).

- Escherichia coli osmotically inducible lipoprotein E (gene osmE).

10 - Escherichia coli peptidoglycan-associated lipoprotein (gene pal).

- Escherichia coli rare lipoproteins A and B (genes rplA and rplB).

- Escherichia coli copper homeostasis protein cutF (or nlpE).

- Escherichia coli plasmids traT proteins.

- Escherichia coli Col plasmids lysis proteins.

15 - A number of Bacillus beta-lactamases.

- Bacillus subtilis periplasmic oligopeptide-binding protein (gene oppA).

- Borrelia burgdorferi outer surface proteins A and B (genes ospA and ospB).

- Borrelia hermsii variable major protein 21 (gene vmp21) and 7 (gene vmp7).

- Chlamydia trachomatis outer membrane protein 3 (gene omp3).

20 - Fibrobacter succinogenes endoglucanase cel-3.

- Haemophilus influenzae proteins Pal and Pcp.

- Klebsiella pullulunase (gene pulA).

- Klebsiella pullulunase secretion protein pulS.

- Mycoplasma hyorhinis protein p37.

25 - Mycoplasma hyorhinis variant surface antigens A, B, and C (genes vlpABC).

- Neisseria outer membrane protein H.8.

- Pseudomonas aeruginosa lipopeptide (gene lppL).

- Pseudomonas solanacearum endoglucanase egl.

- Rhodopseudomonas viridis reaction center cytochrome subunit (gene cytC).

30 - Rickettsia 17 Kd antigen.

- Shigella flexneri invasion plasmid proteins mxiJ and mxiM.

- Streptococcus pneumoniae oligopeptide transport protein A (gene amiA).

- Treponema pallidium 34 Kd antigen.

- Treponema pallidium membrane protein A (gene tmpA).

- Vibrio harveyi chitobiase (gene chb).

- Yersinia virulence plasmid protein yscJ.

- Halocyanin from Natrobacterium pharaonis [4], a membrane associated copper-binding protein. This is the first archaebacterial protein known to be modified in such a fashion).

5   From the precursor sequences of all these proteins, a consensus pattern was derived and a set of rules to identify this type of post-translational modification.

Consensus pattern {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1) The cysteine must be between positions 15 and
10   35 of the sequence in consideration. 2) There must be at least one Lys or one Arg in the first seven positions of the sequence. Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROTsome 100 prokaryotic proteins. Some of them are not membrane lipoproteins, but at least half of them could be.

15   [1] Hayashi S., Wu H.C. J. Bioenerg. Biomembr. 22:451-471(1990).

[2] Klein P., Somorjai R.L., Lau P.C.K. Protein Eng. 2:15-20(1988).

[3] von Heijne G. Protein Eng. 2:531-534(1989).

[4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).

20

901. Ribosome recycling factor (RRF)

The ribosome recycling factor (RRF / ribosome release factor) dissociates the ribosome from the mRNA after termination of translation, and is essential bacterial growth [1]. Thus
25   ribosomes are "recycled" and ready for another round of protein synthesis. Number of members: 27

[1] Janosi L, Shimizu I, Kaji A; Medline: 94240115   Ribosome recycling factor (ribosome releasing factor) is essential for bacterial growth." Proc Natl Acad Sci U S A 1994;91:4249-
30   4253.

902. S-layer homology(SLH)

S-layers are paracrystalline mono-layered assemblies of (glyco)proteins which coat the surface of bacteria [1]. Several S-layer proteins and some other cell wall proteins contain one or more copies of a domain of about 50-60 residues, which has been called SLH (for S-layer homology) [2]. There is strong evidence that this domain serves as an anchor to the

5   peptidoglycan [3]. The SLH domain has been found in:

- S-layer glycoprotein of Acetogenium kivui (3 copies).

- S-layer 125 Kd protein of Bacillus sphaericus (3 copies).

- S-layer protein of Bacillus anthracis (3 copies).

- S-layer protein of Bacillus licheniformis (3 copies).

10   - S-layer protein (HWP) from Bacillus brevis strain HPD31 (3 copies).

- Middle cell wall protein (MWP) from Bacillus brevis strain 47 (3 copies).

- S-layer protein (p100) of Thermus thermophilus (1 copy).

- Outer membrane protein Omp-alpha from Thermotoga maritima (1 copy).

- Cellulosome anchoring protein (gene ancA), outer layer protein B (OlpB) and a further

15   potential cell surface glycoprotein from Clostridium thermocellum (3 copies; the first copy is missing its N-terminal third which is appended to the end of the third copy; may have arisen by circular permutation).

- Amylopullulanase (gene amyB) from Thermoanaerobacter thermosulfurogenes (3 copies)

- Amylopullulanase (gene aapT) from Bacillus strain XAL-601 (3 copies).

20   - Endoglucanase from Bacillus strain KSM-635 (3 copies).

- Exoglucanase (gene xynX) from Clostridium thermocellum (3 copies).

- Xylanase A (gene xynA) from Thermoanaerobacter saccharolyticum (2 copies; 3 copies if a frameshift is taken into account).

- Protein involved in butirosin production (ButB) from Bacillus circulans (2 incomplete

25   copies; 3 copies if three frameshifts are taken into account).

- Two hypothetical proteins from Synechocystis strain PCC 6803 (1 copy each).

- A hypothetical protein with sequence similarity to amylopullulanases found 3' of amylase gene from Bacillus circulans (fragment of 1 copy; 3 copies if two frameshifts are taken into account).

30   SLH domains are found at the N- or C-termini of mature proteins. They occur in single copy followed by a predicted coiled coil domain, or in three contiguous copies. Structurally, the SLH domain is predicted to contain two alpha-helices flanking a beta strand. The SLH sequences are fairly divergent with an average identity of about 25%. It is however possible

to build a sequence pattern that starts at the second position of the domain and that spans 3/4 of its length.

Consensus pattern[LVFYT]-x-[DA]-x(2,5)-[DNGSATPHY]-[FYWPDA]-x(4)-[LIV]-x(2)-
5   [GTALV]-x(4,6)-[LIVFYC]-x(2)-G-x-[PGSTA]-x(2,3)-[MFYA]-x- [PGAV]-x(3,10)-
[LIVMA]-[STKR]-[RY]-x-[EQ]-x-[STALIVM] Sequences known to belong to this class
detected by the pattern ALL. Other sequence(s) detected in SWISS-PROTNONE.

[1] Beveridge T.J. Curr. Opin. Struct. Biol. 4:204-212(1994).
10   [2] Lupas A., Engelhardt H., Peters J., Santarius U., Volker S., Baumeister W. J. Bacteriol.
176:1224-1233(1994).

[3] Lemaire M., Ohayon H., Gounon P., Fujino T., Beguin P. J. Bacteriol. 177:2451-
2459(1995).

15

903. Queuine tRNA-ribosyltransferase (TGT)

This is a family of queuine tRNA-ribosyltransferases EC:2.4.2.29, also known as tRNA-
guanine transglycosylase and guanine insertion enzyme. Queuine tRNA-ribosyltransferase
modifies tRNAs for asparagine, aspartic acid, histidine and tyrosine with queuine. It catalyses
20   the exchange of guanine-34 at the wobble position with 7-aminomethyl-7-deazaguanine, and
the addition of a cyclopentenediol moiety to 7-aminomethyl-7-deazaguanine-34 tRNA;
giving a hypermodified base queuine in the wobble position [1,2].The aligned region contains
a zinc binding motif C-x-C-x2-C-x29-H, and important tRNA and 7-aminomethyl-
7deazaguanine binding residues [1]. Number of members: 27

25

[1] Romier C, Reuter K, Suck D, Ficner R; Medline: 96256303 Crystal structure of tRNA-
guanine transglycosylase: RNA modification by base exchange." EMBO J 1996;15:2850-
2857.

[2] Garcia GA, Koch KA, Chong S; Medline: 93287116 tRNA-guanine transglycosylase
30   from Escherichia coli. Overexpression, purification and quaternary structure." J Mol Biol
1993;231:489-497.

904. ThiC Family (ThiC)

ThiC is found within the thiamine biosynthesis operon. ThiC is involved in pyrimidine biosynthesis [2]. ThiC catalyzes the substitution of the pyrophosphate of 2-methyl-4-amino-5-hydroxymethylpyrimidine pyrophosphate by 4-methyl-5-(beta-hydroxyethyl)thiazole phosphate to yield thiamine phosphate [3]. Number of members: 12

[1] Vander Horn PB, Backstrom AD, Stewart V, Begley TP; Medline: 93163063 Structural genes for thiamine biosynthetic enzymes (thiCEFGH) in Escherichia coli K-12." J Bacteriol 1993;175:982-992.

[2] Begley TP, Downs DM, Ealick SE, McLafferty FW, Van Loon AP, Taylor S, Campobasso N, Chiu HJ, Kinsland C, Reddick JJ, Xi J; Medline: 99311269 Thiamin biosynthesis in prokaryotes." Arch Microbiol 1999;171:293-300.

[3] Zhang Y, Taylor SV, Chiu HJ, Begley TP; Medline: 97284509 Characterization of the Bacillus subtilis thiC operon involved in thiamine biosynthesis." J Bacteriol 1997;179:3030-3035.

905. Putative tRNA binding domain (tRNA_bind)

This domain is found in prokaryotic methionyl-tRNA synthetases, prokaryotic phenylalanyl tRNA synthetases the yeast GU4 nucleic-binding protein (G4p1 or p42, ARC1) [2], human tyrosyl-tRNA synthetase [1], and endothelial-monocyte activating polypeptide II. G4p1 binds specifically to tRNA form a complex with methionyl-tRNA synthetases [2]. In human tyrosyl-tRNA synthetase this domain may direct tRNA to the active site of the enzyme [2]. This domain may perform a common function in tRNA aminoacylation [1]. Number of members: 12

[1] Kleeman TA, Wei D, Simpson KL, First EA; Medline: 97306356 Human tyrosyl-tRNA synthetase shares amino acid sequence homology with a putative cytokine." J Biol Chem 1997;272:14420-14425.

[2] Simos G, Segref A, Fasiolo F, Hellmuth K, Shevchenko A, Mann M, Hurt EC; Medline: 97050848 The yeast protein Arc1p binds to tRNA and functions as a cofactor for the methionyl-and glutamyl-tRNA synthetases." EMBO J 1996;15:5437-5448.

906. UbiA prenyltransferase family signature (UbiA)

The following prenyltransferases are evolutionary related [1,2]:

- Bacterial 4-hydroxybenzoate octaprenyltransferase (gene ubiA).

- Yeast mitochondrial para-hydroxybenzoate--polyprenyltransferase (gene COQ2).

- Protoheme IX farnesyltransferase (heme O synthase) from yeast and mammals (gene

5   COX10) and from bacteria (genes cyoE or ctaB).

These proteins probably contain seven transmembrane segments. The best conserved region is located in a loop between the second and third of these segments and was used as a signature pattern.

10

Consensus pattern N-x(3)-[DE]-x(2)-[LIF]-D-x(2)-[VM]-x-R-[ST]-x(2)-R-x(4)-G Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROTNONE.

15   [1] Melzer M., Heide L. Biochim. Biophys. Acta 1212:93-102(1994).

[2] Mogi T., Saiki K., Anraku Y. Mol. Microbiol. 14:391-398(1994).

907. Uncharacterized protein family UPF0044 signature (UPF0044)

20   The following uncharacterized proteins have been shown [1] to be highly similar:

- Bacillus subtilis hypothetical protein yqeI.

- Escherichia coli hypothetical protein yhbY and HI1333, the corresponding Haemophilus influenzae protein.

- Methanococcus jannaschii hypothetical protein MJ0652.

25   These are small proteins of 10 to 15 Kd. They can be picked up in the database by the following pattern. This pattern is located in the N-terminal part of these proteins.

Consensus pattern L-[ST]-x(3)-K-x(3)-[KR]-[SGA]-x-[GA]-H-x-L-x-P-[LIV]-x(2)- [LIV]-[GA]-x(2)-G Sequences known to belong to this class detected by the patternALL.

30

908. ATP synthase (C/AC39) subunit (vATP-synt_AC39)

This family includes the AC39 subunit from vacuolar ATP synthase Swiss:P32366 [1], and the C subunit from archaebacterial ATP synthase [2]. The family also includes subunit C

from the Sodium transporting ATP synthase from Enterococcus hirae Swiss:P43456 [3]. Number of members: 12

5 [1] Bauerle C, Ho MN, Lindorfer MA, Stevens TH; Medline: 93286119 The Saccharomyces cerevisiae VMA6 gene encodes the 36-kDa subunit of the vacuolar H(+)-ATPase membrane sector." J Biol Chem 1993;268:12749-12757.

[2] Wilms R, Freiberg C, Wegerle E, Meier I, Mayer F, Muller V; Medline: 96324968 Subunit structure and organization of the genes of the A1A0 ATPase from the Archaeon Methanosarcina mazei Go1." J Biol Chem 1996;271:18843-18852.

10 [3] Takase K, Kakinuma S, Yamato I, Konishi K, Igarashi K, Kakinuma Y; Medline: 94209269 Sequencing and characterization of the ntp gene cluster for vacuolar- type Na(+)- translocating ATPase of Enterococcus hirae." J Biol Chem 1994;269:11037-11044.

15 909. ATP synthase (E/31 kDa) subunit (vATP-synt_E)

This family includes the vacuolar ATP synthase E subunit [1], as well as the archaebacterial ATP synthase E subunit [2]. Number of members: 24

[1] Foury F; Medline: 91009356 The 31-kDa polypeptide is an essential subunit of the 20 vacuolar ATPase in Saccharomyces cerevisiae." J Biol Chem 1990;265:18554-18560.

[2] Wilms R, Freiberg C, Wegerle E, Meier I, Mayer F, Muller V; Medline: 96324968 Subunit structure and organization of the genes of the A1A0 ATPase from the Archaeon Methanosarcina mazei Go1." J Biol Chem 1996;271:18843-18852.

25

910. (WW)

The WW domain [1-4,E1] (also known as rsp5 or WWP) has been originally discovered as a short conserved region in a number of unrelated proteins, among them dystrophin, the gene responsible for Duchenne muscular dystrophy. The domain, which spans about 35 residues, is 30 repeated up to 4 times in some proteins. It has been shown [5] to bind proteins with particular proline- motifs, [AP]-P-P-[AP]-Y, and thus resembles somewhat SH3 domains. It appears to contain beta-strands grouped around four conserved aromatic positions; generally Trp. The name WW or WWP derives from the presence of these Trp as well as that of a conserved Pro.

It is frequently associated with other domains typical for proteins in signal transduction processes.

5 Proteins containing the WW domain are listed below.

- Dystrophin, a multidomain cytoskeletal protein. Its longest alternatively spliced form consists of an N-terminal actin-binding domain, followed by 24 spectrin-like repeats, a cysteine-rich calcium-binding domain and a C- terminal globular domain. Dystrophin form tetramers and is thought to have multiple functions including involvement in membrane stability, transduction of contractile forces to the extracellular environment and organization

10 of membrane specialization. Mutations in the dystrophin gene lead to muscular dystrophy of Duchenne or Becker type. Dystrophin contains one WW domain C-terminal of the spectrin-repeats.

- Utrophin, a dystrophin-like protein of unknown function.

- Vertebrate YAP protein is a substrate of an unknown serine kinase. It binds to the SH3

15 domain of the Yes oncoprotein via a proline-rich region. This protein appears in alternatively spliced isoforms, containing either one or two WW domains [6].

- Mouse NEDD-4 plays a role in the embryonic development and differentiation of the central nervous system. It contains 3 WW modules followed by a HECT domain. The human ortholog contains 4 WW domains, but the third WW domain is probably spliced resulting in

20 an alternate NEDD-4 protein with only 3 WW modules [3].

- Yeast RSP5 is similar to NEDD-4 in its molecular organization. It contains an N-terminal C2 domain (see <PDOC00380>, followed by a histidine-rich region, 3 WW domains and a HECT domain.

- Rat FE65, a transcription-factor activator expressed preferentially in liver. The activator

25 domain is located within the N-terminal 232 residues of FE65, which also contain the WW domain.

- Yeast ESS1/PTF1, a putative peptidyl prolyl cis-trans isomerase from family ppiC (see <PDOC00840>). A related protein, dodo (gene dod) exists in Drosophila and in mammals (gene PIN1).

30 - Tobacco DB10 protein. The WW domain is located N-terminal to the region with similarity to ATP-dependent RNA helicases.

- IQGAP, a human GTPase activating protein acting on ras. It contains an N- terminal domain similar to fly muscle mp20 protein and a C-terminal ras GTPase activator domain.

- Yeast pre-mRNA processing protein PRP40, Caenorhabditis elegans ZK1098.1 and fission yeast SpAC13C5.02 are related proteins with similarity to MYO2- type myosin, each containing two WW-domains at the N-terminus.

- Caenorhabditis elegans hypothetical protein C38D4.5, which contains one WW module, a
5 PH domain (see <PDOC50003>) and a C-terminal phosphatidylinositol 3-kinase domain.

- Yeast hypothetical protein YFL010c.

For the sensitive detection of WW domains, a profile was developed which spans the whole homology region as well as a pattern.

10 Consensus pattern W-x(9,11)-[VFY]-[FYW]-x(6,7)-[GSTNE]-[GSTQCR]-[FYW]-x(2)-P

Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT8. Sequences known to belong to this class detected by the profileALL.

15 [1] Bork P., Sudol M. Trends Biochem. Sci. 19:531-533(1994).

[2] Andre B., Springael J.Y. Biochem. Biophys. Res. Commun. 205:1201-1205(1994).

[3] Hofmann K.O., Bucher P. FEBS Lett. 358:153-157(1995).

[4] Sudol M., Chen H.I., Bougeret C., Einbond A., Bork P. FEBS Lett. 369:67-71(1995).

[5] Chen H.I., Sudol M. Proc. Natl. Acad. Sci. U.S.A. 92:7819-7823(1995).

20 [6] Sudol M., Bork P., Einbond A., Kastury K., Druck T., Negrini M., Huebner K., Lehman D. J. Biol. Chem. 270:14733-14741(1995).

911. Xeroderma pigmentosum (XP) [1] (XPG_1)

25 Xeroderma pigmentosum (XP) [1] is a human autosomal recessive disease, characterized by a high incidence of sunlight-induced skin cancer. People's skin cells with this condition are hypersensitive to ultraviolet light, due to defects in the incision step of DNA excision repair. There are a minimum of seven genetic complementation groups involved in this pathway: XP-A to XP-G. The defect in XP-G can be corrected by a 133 Kd nuclear protein called XPG
30 (or XPGC) [2].

XPG belongs to a family of proteins [2,3,4,5,6] that are composed of two main subsets:

- Subset 1, to which belongs XPG, RAD2 from budding yeast and rad13 from fission yeast. RAD2 and XPG are single-stranded DNA endonucleases [7,8]. XPG makes the 3'incision in human DNA nucleotide excision repair [9].

- Subset 2, to which belongs mouse and human FEN-1, rad2 from fission yeast, and RAD27
5 from budding yeast. FEN-1 is a structure-specific endonuclease.

In addition to the proteins listed in the above groups, this family also includes:

- Fission yeast exo1, a 5'->3' double-stranded DNA exonuclease that could act in a pathway that corrects mismatched base pairs.

10 - Yeast EXO1 (DHS1), a protein with probably the same function as exo1.

- Yeast DIN7.

Sequence alignment of this family of proteins reveals that similarities are largely confined to two regions. The first is located at the N-terminal extremity (N-region) and corresponds to
15 the first 95 to 105 amino acids. The second region is internal (I-region) and found towards the C-terminus; it spans about 140 residues and contains a highly conserved core of 27 amino acids that includes a conserved pentapeptide (E-A-[DE]-A-[QS]). It is possible that the conserved acidic residues are involved in the catalytic mechanism of DNA excision repair in XPG. The amino acids linking the N- and I-regions are not conserved; indeed, they are
20 largely absent from proteins belonging to the second subset.

Two signature patterns were developed for these proteins. The first corresponds to the central part of the N-region, the second to part of the I-region and includes the putative catalytic core pentapeptide.
25

Consensus pattern [VI]-[KRE]-P-x-[FYIL]-V-F-D-G-x(2)-[PIL]-x-[LVC]-K Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROTNONE.

30 Consensus pattern [GS]-[LIVM]-[PER]-[FYS]-[LIVM]-x-A-P-x-E-A-[DE]-[PAS]- [QS]-[CLM] Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROTNONE.

[1] Tanaka K., Wood R.D. Trends Biochem. Sci. 19:83-86(1994).

[2] Scherly D., Nouspikel T., Corlet J., Ucla C., Bairoch A., Clarkson S.G. Nature 363:182-185(1993).

[3] Carr A.M., Sheldrick K.S., Murray J.M., Al-Harithy R., Watts F.Z., Lehmann A.R. Nucleic Acids Res. 21:1345-1349(1993).

[4] Murray J.M., Tavassoli M., Al-Harithy R., Sheldrick K.S., Lehmann A.R., Carr A.M., Watts F.Z. Mol. Cell. Biol. 14:4878-4888(1994).

[5] Harrington J.J., Lieber M.R. Genes Dev. 8:1344-1355(1994).

[6] Szankasi P., Smith G.R. Science 267:1166-1169(1995).

[7] Habraken Y., Sung P., Prakash L., Prakash S. Nature 366:365-368(1993).

[8] O'Donovan A., Scherly D., Clarkson S.G., Wood R.D. J. Biol. Chem. 269:15965-15968(1994).

[9] O'Donovan A., Davies A.A., Moggs J.G., West S.C., Wood R.D. Nature 371:432-435(1994).

912. 5-formyltetrahydrofolate cyclo-ligase (5-FTHF_cyc-lig)

5-formyltetrahydrofolate cyclo-ligase or methenyl-THF synthetase EC:6.3.3.2 catalyses the interchange of 5-formyltetrahydrofolate (5-FTHF) to 5-10-methenyltetrahydrofolate, this requires ATP and Mg2+ [1]. 5-FTHF is used in chemotherapy where it is clinically known as Leucovorin [2].

Number of members: 23

[1] Dayan A, Bertrand R, Beauchemin M, Chahla D, Mamo A, Filion M, Skup D, Massie B, Jolivet J; Medline: 96096540 Cloning and characterization of the human 5,10-methenyltetrahydrofolate synthetase-encoding cDNA." Gene 1995;165:307-311.
[2] Maras B, Stover P, Valiante S, Barra D, Schirch V; Medline: 94308074 Primary structure and tetrahydropteroylglutamate binding site of rabbit liver cytosolic 5,10-methenyltetrahydrofolate synthetase." J Biol Chem 1994;269:18429-18433.

913. Cytosolic long-chain acyl-CoA thioester hydrolase (Acyl-CoA_hydro)

This family consist of various cytosolic long-chain acyl-CoA thioester hydrolases including human and rat [1,2]. The aligned region is repeated with in the sequence of human and rat

cytosolic long-chain acyl-CoA thioester hydrolases of this family. Long-chain acyl-CoA hydrolases hydrolyse palmitoyl-CoA to CoA and palmitate, they also catalyse the hydrolysis of other long chain fatty acyl-CoA thioesters. Long-chain acyl-CoA hydrolases are present in all living organisms and they may provide a mechanism for the control of lipid metabolism [1].

Number of members: 24

[1]Yamada J, Furihata T, Iida N, Watanabe T, Hosokawa M, Satoh T, Someya A, Nagaoka I, Suga T; Medline: 97236308   Molecular cloning and expression of cDNAs encoding rat brain and liver cytosolic long-chain acyl-CoA hydrolases." Biochem Biophys Res Commun 1997;232:198-203.

[2] Broustas CG, Larkins LK, Uhler MD, Hajra AK; Medline: 96209964   Molecular cloning and expression of cDNA encoding rat brain cytosolic acyl-coenzyme A thioester hydrolase." J Biol Chem 1996;271:10470-10476.

914. Agglutinin

Lectin (probable mannose binding)

Members of this family are plant lectins. Many if not all are mannose specific.

Number of members: 87

[1] Wright CS, Hester G; Medline: 97094989   The 2.0 A structure of a cross-linked complex between snowdrop lectin and a branched mannopentaose: evidence for two unique binding modes." Structure 1996;4:1339-1352.

915. (ANF_RECEPTORS)

Natriuretic peptides are hormones involved in the regulation of fluid and electrolyte homeostasis. These hormones stimulate the intracellular production of cyclic GMP as a second messenger.

Currently, three types of natriuretic peptide receptors are known [1,2]. Two express guanylate cyclase activity: GC-A (or ANP-A) which seems specific to atrial natriuretic peptide (ANP), and GC-B (or ANP-B) which seems to be stimulated more effectively by brain natriuretic

peptide (BNP) than by ANP. The third receptor (ANP-C) is probably responsible for the clearance of ANP from the circulation and does not play a role in signal transduction.

5   GC-A and GC-B are plasma membrane-bound proteins that share the following topology: an N-terminal extracellular domain which acts as the ligand binding region, then a transmembrane domain followed by a large cytoplasmic C- terminal region that can be subdivided into two domains: a protein kinase-like domain (see <PDOC00100>) that appears important for proper signalling and a guanylate cyclase catalytic domain (see <PDOC00425>). The topology of ANP-C is different: like GC-A and -B it possesses an

10   extracellular ligand-binding region and a transmembrane domain, but its cytoplasmic domain is very short.

A pattern was developed from the ligand-binding region of natriuretic peptide receptors based on a highly conserved region located in the N-terminal part of the domain.

15

Consensus patternG-P-x-C-x-Y-x-A-A-x-V-x-R-x(3)-H-W Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROTNONE.

[1] Garbers D.L. New Biol. 2:499-504(1990).

20   [2] Schulz S., Chinkers M., Garbers D.L. FASEB J. 2:2026-2035(1989).

916. (Apocytochrome)
Cytochrome c family heme-binding site signature

25 · In proteins belonging to cytochrome c family [1], the heme group is covalently attached by thioether bonds to two conserved cysteine residues. The consensus sequence for this site is Cys-X-X-Cys-His and the histidine residue is one of the two axial ligands of the heme iron. This arrangement is shared by all proteins known to belong to cytochrome c family, which presently includes cytochromes c, c', c1 to c6, c550 to c556, cc3/Hmc, cytochrome f and

30   reaction center cytochrome c.

Consensus patternC-{CPWHF}-{CPWR}-C-H-{CFYW} Sequences known to belong to this class detected by the patternALL, except for four cytochrome c's which lack the first thioether bond. Other sequence(s) detected in SWISS-PROT454.

Note: some cytochrome c's have more than a single bound heme groupc4 has 2, c7 has 3, c3 has 4, the reaction center has 4, and cc3/Hmc has 16 !

5   [1] Mathews F.S. Prog. Biophys. Mol. Biol. 45:1-56(1985).

917. ATP-synt_A-c. ATP synthase Alpha chain, C terminal

[1] Medline: 94344236. Structure at 2.8 A resolution of F1-ATPase from bovine heart mitochondria. Abrahams JP, Leslie AG, Lutter R, Walker JE; Nature 1994;370:621-628.

10   Number of members: 125

918. (Basic)

Myc-type, 'helix-loop-helix' dimerization domain signature

HELIX_LOOP_HELIX

15

A number of eukaryotic proteins, which probably are sequence specific DNA- binding proteins that act as transcription factors, share a conserved domain of 40 to 50 amino acid residues. It has been proposed [1] that this domain is formed of two amphipathic helices joined by a variable length linker region that could form a loop. This 'helix-loop-helix' (HLH)

20   domain mediates protein dimerization and has been found in the proteins listed below [2,3,E1,E2]. Most of these proteins have an extra basic region of about 15 amino acid residues that is adjacent to the HLH domain and specifically binds to DNA. They are refered as basic helix-loop-helix proteins (bHLH), and are classified in two groups: class A (ubiquitous) and class B (tissue-specific). Members of the bHLH family bind variations on

25   the core sequence 'CANNTG', also refered to as the E-box motif. The homo- or heterodimerization mediated by the HLH domain is independent of, but necessary for DNA binding, as two basic regions are required for DNA binding activity. The HLH proteins lacking the basic domain (Emc, Id) function as negative regulators since they form heterodimers, but fail to bind DNA. The hairy-related proteins (hairy, E(spl), deadpan) also

30   repress transcription although they can bind DNA. The proteins of this subfamily act together with co-repressor proteins, like groucho, through their C-terminal motif WRPW.
- The myc family of cellular oncogenes [4], which is currently known to contain four members: c-myc [E3], N-myc, L-myc, and B-myc. The myc genes are thought to play a role in cellular differentiation and proliferation.

- Proteins involved in myogenesis (the induction of muscle cells). In mammals MyoD1 (Myf-3), myogenin (Myf-4), Myf-5, and Myf-6 (Mrf4 or herculin), in birds CMD1 (QMF-1), in Xenopus MyoD and MF25, in Caenorhabditis elegans CeMyoD, and in Drosophila nautilus (nau).

5   - Vertebrate proteins that bind specific DNA sequences ('E boxes') in various immunoglobulin chains enhancers: E2A or ITF-1 (E12/pan-2 and E47/pan-1), ITF-2 (tcf4), TFE3, and TFEB.

- Vertebrate neurogenic differentiation factor 1 that acts as differentiation factor during neurogenesis.

10   - Vertebrate MAX protein, a transcription regulator that forms a sequence- specific DNA-binding protein complex with myc or mad.

- Vertebrate Max Interacting Protein 1 (MXI1 protein) which acts as a transcriptional repressor and may antagonize myc transcriptional activity by competing for max.

- Proteins of the bHLH/PAS superfamily which are transcriptional activators. In mammals,

15   AH receptor nuclear translocator (ARNT), single-minded homologs (SIM1 and SIM2), hypoxia-inducible factor 1 alpha (HIF1A), AH receptor (AHR), neuronal pas domain proteins (NPAS1 and NPAS2), endothelial pas domain protein 1 (EPAS1), mouse ARNT2, and human BMAL1. In drosophila, single-minded (SIM), AH receptor nuclear translocator (ARNT), trachealess protein (TRH), and similar protein (SIMA).

20   - Mammalian transcription factors HES, which repress transcription by acting on two types of DNA sequences, the E box and the N box.

- Mammalian MAD protein (max dimerizer) which acts as transcriptional repressor and may antagonize myc transcriptional activity by competing for max.

- Mammalian Upstream Stimulatory Factor 1 and 2 (USF1 and USF2), which bind to a

25   symmetrical DNA sequence that is found in a variety of viral and cellular promoters.

- Human lyl-1 protein; which is involved, by chromosomal translocation, in T- cell leukemia.

- Human transcription factor AP-4.

- Mouse helix-loop-helix proteins MATH-1 and MATH-2 which activate E box- dependent transcription in collaboration with E47.

30   - Mammalian stem cell protein (SCL) (also known as tal1), a protein which may play an important role in hemopoietic differentiation. SCL is involved, by chromosomal translocation, in stem-cell leukemia.

- Mammalian proteins Id1 to Id4 [5]. Id (inhibitor of DNA binding) proteins lack a basic DNA-binding domain but are able to form heterodimers with other HLH proteins, thereby inhibiting binding to DNA.

- Drosophila extra-macrochaetae (emc) protein, which participates in sensory organ
5   patterning by antagonizing the neurogenic activity of the achaete- scute complex. Emc is the homolog of mammalian Id proteins.

- Human Sterol Regulatory Element Binding Protein 1 (SREBP-1), a transcriptional activator that binds to the sterol regulatory element 1 (SRE-1) found in the flanking region of the LDLR gene and in other genes.

10   - Drosophila achaete-scute (AS-C) complex proteins T3 (l'sc), T4 (scute), T5 (achaete) and T8 (asense). The AS-C proteins are involved in the determination of the neuronal precursors in the peripheral nervous system and the central nervous system.

- Mammalian homologs of achaete-scute proteins, the MASH-1 and MASH-2 proteins.

- Drosophila atonal protein (ato) which is involved in neurogenesis.

15   - Drosophila daughterless (da) protein, which is essential for neurogenesis and sex-determination.

- Drosophila deadpan (dpn), a hairy-like protein involved in the functional differentiation of neurons.

- Drosophila delilah (dei) protein, which is plays an important role in the differentiation of
20   epidermal cells into muscle.

- Drosophila hairy (h) protein, a transcriptional repressor which regulates the embryonic segmentation and adult bristle patterning.

- Drosophila enhancer of split proteins E(spl), that are hairy-like proteins active during neurogenesis. also act as transcriptional repressors.

25   - Drosophila twist (twi) protein, which is involved in the establishment of germ layers in embryos.

- Maize anthocyanin regulatory proteins R-S and LC.

- Yeast centromere-binding protein 1 (CPF1 or CBF1). This protein is involved in chromosomal segregation. It binds to a highly conserved DNA sequence, found in centromers
30   and in several promoters.

- Yeast INO2 and INO4 proteins.

- Yeast phosphate system positive regulatory protein PHO4 which interacts with the upstream activating sequence of several acid phosphatase genes.

- Yeast serine-rich protein TYE7 that is required for ty-mediated ADH2 expression.

- Neurospora crassa nuc-1, a protein that activates the transcription of structural genes for phosphorus acquisition.

- Fission yeast protein esc1 which is involved in the sexual differentiation process.

5 The schematic representation of the helix-loop-helix domain is shown here:

xxxxxxxxxxxxxxxxxxxxxxxx------------------xxxxxxxxxxxxxxxxxxxxxxxx Amphipathic helix 1 Loop Amphipathic helix 2

The signature pattern that had been developed to detect this domain spans completely the
10 second amphipathic helix.

Consensus pattern[DENSTAP]-[KR]-[LIVMAGSNT]-{FYWCPHKR}-[LIVMT]-[LIVM]-x(2)-[STAV]-[LIVMSTACKR]-x-[VMFYH]-[LIVMTA]-{P}-{P}- [LIVMRKHQ]

Sequences known to belong to this class detected by the pattern the majority but far from all.
15 Other sequence(s) detected in SWISS-PROT135.

[1] Murre C., McCaw P.S., Baltimore D. Cell 56:777-783(1989).

[2] Garrel J., Campuzano S. BioEssays 13:493-498(1991).

[3] Kato G.J., Dang C.V. FASEB J. 6:3065-3072(1992).

20 [4] Krause M., Fire A., Harrison S.W., Priess J., Weintraub H. Cell 63:907-919(1990).

[5] Riechmann V., van Cruechten I., Sablitzky F. Nucleic Acids Res. 22:749-755(1994).

919. (Beta-lactamase)

Beta-lactamases classes -A, -C, and -D active site

25

Beta-lactamases (EC 3.5.2.6) [1,2] are enzymes which catalyze the hydrolysis of an amide bond in the beta-lactam ring of antibiotics belonging to the penicillin/cephalosporin family. Four kinds of beta-lactamase have been identified [3]. Class-B enzymes are zinc containing proteins whilst class -A, C and D enzymes are serine hydrolases. The three
30 classes of serine beta-

lactamases are evolutionary related and belong to a superfamily [4] that also includes DD-peptidases and a variety of other penicillin-binding proteins (PBP's). All these proteins contain a Ser-x-x-Lys motif, where the serine is the active site residue. Although clearly homologous, the sequences of the three classes of serine beta-lactamases exhibit a large

728

degree of variability and only a small number of residues are conserved in addition to the catalytic serine.

Since a pattern detecting all serine beta-lactamases would also pick up many unrelated
5   sequences, it was decided to provide specific patterns, centered on the active site serine, for each of the three classes.

Consensus pattern [FY]-x-[LIVMFY]-x-S-[TV]-x-K-x(4)-[AGLM]-x(2)-[LC] [S is the active site residue] Sequences known to belong to this class detected by the patternALL class-A
10   beta-lactamases. Other sequence(s) detected in SWISS-PROT7.

Consensus pattern F-E-[LIVM]-G-S-[LIVMG]-[SA]-K [The first S is the active site residue] Sequences known to belong to this class detected by the patternALL class-C beta-lactamases. Other sequence(s) detected in SWISS-PROTNONE.
15

Consensus pattern [PA]-x-S-[ST]-F-K-[LIV]-[PAL]-x-[STA]-[LI] [S is the active site residue] Sequences known to belong to this class detected by the patternALL class-D beta-lactamases. Other sequence(s) detected in SWISS-PROTNONE.

20   [1] Ambler R.P. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 289:321-331(1980).

[2] Pastor N., Pinero D., Valdes A.M., Soberon X. Mol. Microbiol. 4:1957-1965(1990).

[3] Bush K. Antimicrob. Agents Chemother. 33:259-263(1989).

[4] Joris B., Ghuysen J.-M., Dive G., Renard A., Dideberg O., Charlier P., Frere J.M., Kelly J.A., Boyington J.C., Moews P.C., Knox J.R. Biochem. J. 250:313-324(1988).
25

920. Biotin protein ligase (BPL)

Biotin is covalently attached at the active site of certain enzymes that transfer carbon dioxide from bicarbonate to organic acids to form cellular metabolites. Biotin protein ligase (BPL) is
30   the enzyme responsible for attaching biotin to a specific lysine at the active site of biotin enzymes. Each organism probably has only one BPL. Biotin attachment is a two step reaction that results in the formation of an amide linkage between the carboxyl group of biotin and the epsilon-amino group of the modified lysine [2].

Number of members: 26

[1] Wilson KP, Shewchuk LM, Brennan RG, Otsuka AJ, Matthews BW; Medline: 93028443

Escherichia coli biotin holoenzyme synthetase/bio repressor crystal structure delineates the

biotin- and DNA-binding domains." Proc Natl Acad Sci USA 1992;89:9257-9261.

5   [2] Chapman-Smith A, Cronan JE Jr; Medline: 10470036   The enzymatic biotinylation of

proteins: a post-translational modification of exceptional specificity." Trends Biochem Sci

1999;24:359-363.

921. (BRCA2_repeat)

10

The alignment covers only the most conserved region of the repeat. Respiratory-chain NADH

dehydrogenase 30 Kd subunit signature

[1] Bork P, Blomberg N, Nilges M; Medline: 96241568   Internal repeats in the BRCA2

15   protein sequence." Nat Genet 1996;13:22-23.

Number of members: 63

922. (C6)

20

This domain of unknown function is found in the C. elegans protein Swiss:Q19522. It is

presumed to be an extracellular domain. The C6 domain contains six conserved cysteine

residues in most copies of the domain. However some copies of the domain are missing

cysteine residues 1 and 3 suggesting that these form a disulphide bridge.

25   Number of members: 23

923. Cadherin cytoplasmic region (Cadherin_C_term)

Cadherins are vital in cell-cell adhesion during tissue differentiation. Cadherins are linked to

30   the cytoskeleton by catenins. Catenins bind to the cytoplasmic tail of the cadherin. Cadherins

cluster to form foci of homophilic binding units. A key determinant to the strength of the

binding that it is mediated by cadherins is the juxtamembrane region of the cadherin. This

region induces clustering and also binds to the protein p120ctn [1].

Number of members: 59

[1] Yap AS, Niessen CM, Gumbiner BM; Medline: 98234411 The juxtamembrane region of the cadherin cytoplasmic tail supports lateral clustering, adhesive strengthening, and interaction with p120ctn." J Cell Biol 1998;141:779-789.

[2] Barth AI, Nathke IS, Nelson WJ; Medline: 97471931 Cadherins, catenins and APC protein: interplay between cytoskeletal complexes and signaling pathways." Curr Opin Cell Biol 1997;9:683-690.

[3] Braga VM, Machesky LM, Hall A, Hotchin NA; Medline: 97327766 The small GTPases Rho and Rac are required for the establishment of cadherin-dependent cell-cell contacts." J Cell Biol 1997;137:1421-1431.

924. Clathrin propeller repeat (Clathrin_propel)

Clathrin is the scaffold protein of the basket-like coat that surrounds coated vesicles. The soluble assembly unit, a triskelion, contains three heavy chains and three light chains in an extended three-legged structure. Each leg contains one heavy and one light chain. The N-terminus of the heavy chain is known as the globular domain, and is composed of seven repeats which form a beta propeller [1].

Number of members: 61

[1] ter Haar E, Musacchio A, Harrison SC, Kirchhausen T; Medline: 99043510 Atomic structure of clathrin: a beta propeller terminal domain joins an alpha zigzag linker." Cell. 1998;95:563-573.

925. Respiratory-chain NADH dehydrogenase 30 Kd subunit signature (complex1_30Kd)

Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex there is one with a molecular weight of 30 Kd (in mammals) which has been found to be:

- Nuclear encoded, as a precursor form with a transit peptide in mammals, and in Neurospora crassa.

- Mitochondrial encoded in Paramecium (protein P1), and in the slime mold Dictyostelium discoideum (ORF 209).

- Chloroplast encoded in various higher plants (ORF 159). It is also present in bacteria:

- In the cyanobacteria Synechocystis strain PCC 6803 (gene ndhJ).

5   - Subunit C of Escherichia coli NADH-ubiquinone oxidoreductase (gene nuoC).

- Subunit NQO5 of Paracoccus denitrificans NADH-ubiquinone oxidoreductase.

This protein, in its mature form, consists of from 157 to 266 amino acid residues. The best conserved region is located in the C-terminal section and can be used as a signature pattern.

10

Consensus pattern E-R-E-x(2)-[DE]-[LIVMFY](2)-x(6)-[HK]-x(3)-[KRP]-x-[LIVM]-[LIVMYS] Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROTNONE.

15   [1] Ragan C.I. Curr. Top. Bioenerg. 15:1-36(1987).

[2]Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).

926. Respiratory-chain NADH dehydrogenase 49 Kd subunit signature (complex1_49Kd)

20   Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex there is one with a molecular weight of 49 Kd

25   (in mammals), which is the third largest subunit of complex I and is a component of the iron-sulfur (IP) fragment of the enzyme. It seems to bind a 4Fe-4S iron-sulfur cluster. The 49 Kd subunit has been found to be:

- Nuclear encoded, as a precursor form with a transit peptide in mammals, and in Neurospora crassa.

30   - Mitochondrial encoded in protozoan such as Paramecium (ORF 400), Leishmania and Trypanosoma (MURF 3).

- Chloroplast encoded in various higher plants (ORF 392).

The 49 Kd subunit is highly similar to [3,4]:

- Subunit D of Escherichia coli NADH-ubiquinone oxidoreductase (gene nuoD).

- Subunit NQO4 of Paracoccus denitrificans NADH-ubiquinone oxidoreductase.

- Subunit 5 of Escherichia coli formate hydrogenlyase (gene hycE).

- Subunit G of Escherichia coli hydrogenase-4 (gene hyfG).

A highly conserved region was seleceted as signature pattern, located in the N-terminal

5   section of this subunit.

Consensus pattern [LIVMH]-H-[RT]-[GA]-x-E-K-[LIVMTN]-x-E-x-[KRQ] Sequences known to belong to this class detected by the patternALL.

10   [1] Ragan C.I. Curr. Top. Bioenerg. 15:1-36(1987).

[2] Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).

[3] Fearnley I.M., Walker J.E. Biochim. Biophys. Acta 1140:105-134(1992).

[4] Weidner U., Geier S., Ptock A., Friedrich T., Leif H., Weiss H. J. Mol. Biol. 233:109-122(1993).

15

927. (COX2)

Cytochrome c oxidase (EC 1.9.3.1) [1,2] is an oligomeric enzymatic complex which is a component of the respiratory chain and is involved in the transfer of electrons from

20   cytochrome c to oxygen. In eukaryotes this enzyme complex is located in the mitochondrial inner membrane; in aerobic prokaryotes it is found in the plasma membrane. The enzyme complex consists of 3-4 subunits (prokaryotes) to up to 13 polypeptides (mammals).

Subunit 2 (CO II) transfers the electrons from cytochrome c to the catalytic subunit 1. It

25   contains two adjacent transmembrane regions in its N-terminus and the major part of the protein is exposed to the periplasmic or to the mitochondrial intermembrane space, respectively. CO II provides the substrate- binding site and contains a copper center called Cu(A), probably the primary acceptor in cytochrome c oxidase. An exception is the corresponding subunit of the cbb3-type oxidase which lacks the copper A redox-center.

30   Several bacterial CO II have a C-terminal extension that contains a covalently bound heme c.

It has been shown [3,4] that nitrous oxide reductase (EC 1.7.99.6) (gene nosZ) of Pseudomonas has sequence similarity in its C-terminus to CO II. This enzyme is part of the bacterial respiratory system which is activated under anaerobic conditions in the presence of

nitrate or nitrous oxide. NosZ is a periplasmic homodimer that contains a dinuclear copper center, probably located in a 3- dimensional fold similar to the cupredoxin-like fold that has been suggested for the copper-binding site of CO II [3].

5   The dinuclear purple copper center is formed by 2 histidines and 2 cysteines [5]. This region was used as a signature pattern. The conserved valine and the conserved methionine are said to be involved in stabilizing the copper-binding fold by interacting with each other.

Consensus pattern V-x-H-x(33,40)-C-x(3)-C-x(3)-H-x(2)-M [The two C's and two H's are
10   copper ligands] Sequences known to belong to this class detected by the patternALL, except for Paramecium primaurelia as well as in some plants where the pattern ends with Thr; an RNA editing event at this position could change this Thr to Met.

Note: cytochrome cbb(3) subunit 2 does not belong to this family.
15

[1] Capaldi R.A., Malatesta F., Darley-Usmar V.M. Biochim. Biophys. Acta 726:135-148(1983).

[2] Garcia-Horsman J.A., Barquera B., Rumbley J., Ma J., Gennis R.B. J. Bacteriol. 176:5587-5600(1994).

20   [3] van der Oost J., Lappalainen P., Musacchio A., Warne A., Lemieux L., Rumbley J., Gennis R.B., Aasa R., Pascher T., Malmstrom B.G., Saraste M. EMBO J. 11:3209-3217(1992).

[4] Zumft W.G., Dreutsch A., Loechelt S., Cuypers H., Friedrich B., Schneider B. Eur. J. Biochem. 208:31-40(1992).
25

928. Cytochrome C assembly protein (CytC_asm)

This family consists of various proteins involved in cytochrome c assembly from mitochondria and bacteria; CycK from Rhizobium[3], CcmC from E. coli and Paracoccus
30   denitrificans [2,1] and orf240 from wheat mitochondria [4]. The members of this family are probably integral membrane proteins with six predicted transmembrane helices. It has been proposed that members of this family comprise a membrane component of an ABC (ATP binding cassette) transporter complex. It is also proposed that this transporter is necessary for transport of some component needed for cytochrome c assembly. One member CycK

contains a putative heme-binding motif [3], orf240 also contains a putative heme-binding motif and is a proposed ABC transporter with c-type heme as its proposed substrate [4]. However it seems unlikely that all members of this family transport heme nor c-type apocytochromes because CcmC in the putative CcmABC transporter transports neither [1].

5   Number of members: 67

[1] Page D, Pearce DA, Norris HA, Ferguson SJ; Medline: 97195802   The Paracoccus denitrificans ccmA, B and C genes: cloning and sequencing, and analysis of the potential of their products to form a haem or apo-c-type cytochrome transporter. MICROBIOLOGY

10   1997;143:563-576.

[2] Thoeny-meyer L, Fischer F, Kunzler P, Ritz D, Hennecke H; Medline: 95362656
   Escherichia coli genes required for cytochrome c maturation." J. BACTERIOL
   1995;177:4321-4326.

[3] Delgado MJ, Yeoman KH, Wu G, Vargas C, Davies A, Poole RK, Johnston AWB,

15   Downie JA; Medline: 95394794   Characterization of the cycHJKL genes involved in cytochrome c biogenesis and symbiotic nitrogen fixation in Rhizobium leguminosarum." J. BACTERIOL 1995;177:4927-4934.

[4] Bonnard G, Grienenberger JM; Medline: 95124303   A gene proposed to encode a transmembrane domain of an ABC transporter is expressed in wheat mitochondria." MOL.

20   GEN. GENET 1995;246:91-99.

929. Cytochrome b559 subunits heme-binding site signature (cytochr_b559)

Cytochrome b559 [1] is an essential component of photosystem II complex from oxygenic

25   photosynthetic organisms. It is an integral thylakoid membrane protein composed of two subunits, alpha (gene psbE) and beta (gene psbF), each of which contains a histidine residue located in a transmembrane region. The two histidines coordinate the heme iron of cytochrome b559.

30   The region around the heme-binding residue of both subunits is very similar and can be used as a signature pattern.

735

Consensus pattern[LIV]-x-[ST]-[LIVF]-R-[FYW]-x(2)-[IV]-H-[STGA]-[LIV]- [STGA]-[IV]-P [H is the heme iron ligand] Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROTNONE.

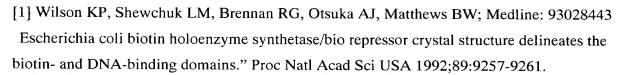5   [1] Pakrasi H.B., de Ciechi P., Whitmarsh J. EMBO J. 10:1619-1627(1991).

930. Cytochrome b/b6 signatures (Cytochrome_b)

10   In the mitochondrion of eukaryotes and in aerobic prokaryotes, cytochrome b is a component of respiratory chain complex III (EC 1.10.2.2) - also known as the bc1 complex or ubiquinol-cytochrome c reductase. In plant chloroplasts and cyanobacteria, there is a analogous protein, cytochrome b6, a component of the plastoquinone-plastocyanin reductase (EC 1.10.99.1), also known as the b6f complex.

15

Cytochrome b/b6 [1,2] is an integral membrane protein of approximately 400 amino acid residues that probably has 8 transmembrane segments. In plants and cyanobacteria, cytochrome b6 consists of two subunits encoded by the petB and petD genes. The sequence of petB is colinear with the N-terminal part of mitochondrial cytochrome b, while petD

20   corresponds to the C-terminal part. Cytochrome b/b6 non-covalently binds two heme groups, known as b562 and b566. Four conserved histidine residues are postulated to be the ligands of the iron atoms of these two heme groups.

Apart from regions around some of the histidine heme ligands, there are a few conserved

25   regions in the sequence of b/b6. The best conserved of these regions includes an invariant P-E-W triplet which lies in the loop that separates the fifth and sixth transmembrane segments. It seems to be important for electron transfer at the ubiquinone redox site - called Qz or Qo (where o stands for outside) - located on the outer side of the membrane.

30   A schematic representation of the structure of cytochrome b/b6 is shown below.

+---Fe-b562----+ | +---Fe-b566--|-+ | | | |
xxxxxxxxxxxHxHxxxxxxxxxxxxHxHxxxxxxxxxxPEWxxxxxxxxxxxxxxxxxxx <--------------

736

---Cytochrome-b-------------------------> <----Cytochrome-b6-petB---------><--Cytochrome-b6-petD----->

Two signature patterns were developed for cytochrome b/b6. The first includes the first
conserved histidine of b/b6, which is a heme b562 ligand; the second includes the conserved
PEW triplet.

Consensus pattern [DENQ]-x(3)-G-[FYWMQ]-x-[LIVMF]-R-x(2)-H [H is a heme b562
ligand] Sequences known to belong to this class detected by the patternALL, except for 5
sequences.

Consensus pattern P-[DE]-W-[FY]-[LFY](2) Sequences known to belong to this class
detected by the patternALL, except for Odocoileus hemionus (mule deer) and Paramecium
tetraurelia cytochrome b.

[1] Howell N. J. Mol. Evol. 29:157-169(1989).

[2] Esposti M.D., de Vries S., Crimi M., Ghelli A., Patarnello T., Meyer A. Biochim.
Biophys. Acta 1143:243-271(1993).

931. Phorbol esters / diacylglycerol binding domain (DAG_PE-bind)

Diacylglycerol (DAG) is an important second messenger. Phorbol esters (PE) are analogues
of DAG and potent tumor promoters that cause a variety of physiological changes when
administered to both cells and tissues. DAG activates a family of serine/threonine protein
kinases, collectively known as protein kinase C (PKC) [1]. Phorbol esters can directly
stimulate PKC. The N- terminal region of PKC, known as C1, has been shown [2] to bind PE
and DAG in a phospholipid and zinc-dependent fashion. The C1 region contains one or two
copies (depending on the isozyme of PKC) of a cysteine-rich domain about 50 amino-acid
residues long and essential for DAG/PE-binding. Such a domain has also been found in the
following proteins:

- Diacylglycerol kinase (EC 2.7.1.107) (DGK) [3], the enzyme that converts DAG into
phosphatidate. It contains two copies of the DAG/PE-binding domain in its N-terminal
section. At least five different forms of DGK are known in mammals.

- N-chimaerin. A brain specific protein which shows sequence similarities with the BCR protein at its C-terminal part and contains a single copy of the DAG/PE-binding domain at its N-terminal part. It has been shown [4,5] to be able to bind phorbol esters.

- The raf/mil family of serine/threonine protein kinases. These protein kinases contain a

5   single N-terminal copy of the DAG/PE-binding domain.

- The unc-13 protein from Caenorhabditis elegans. Its function is not known but it contains a copy of the DAG/PE-binding domain in its central section and has been shown to bind specifically to a phorbol ester in the presence of calcium [6].

- The vav oncogene. Vav was generated by a genetic rearrangement during gene transfer

10   assays. Its expression seems to be restricted to cells of hematopoeitic origin. Vav seems [5,7] to contain a DAG/PE-binding domain in the central part of the protein.

- The Drosophila GTPase activating protein rotund.

The DAG/PE-binding domain binds two zinc ions; the ligands of these metal ions are

15   probably the six cysteines and two histidines that are conserved in this domain. A signature pattern was developed that spans completely the DAG/PE domain.

Consensus pattern H-x-[LIVMFYW]-x(8,11)-C-x(2)-C-x(3)-[LIVMFC]-x(5,10)- C-x(2)-C-x(4)-[HD]-x(2)-C-x(5,9)-C [All the C and H are involved in binding Zinc] Sequences known

20   to belong to this class detected by the pattern ALL, except a few DGK's.

[1] Azzi A., Boscoboinik D., Hensey C. Eur. J. Biochem. 208:547-557(1992).

[2] Ono Y., Fujii T., Igarashi K., Kuno T., Tanaka C, Kikkawa U., Nishizuka Y. Proc. Natl. Acad. Sci. U.S.A. 86:4868-4871(1989).

25   [3] Sakane F., Yamada K., Kanoh H., Yokoyama C., Tanabe T. Nature 344:345-348(1990).

[4] Ahmed S., Kozma R., Monfries C., Hall C., Lim H.H., Smith P., Lim L. Biochem. J. 272:767-773(1990).

[5] Ahmed S., Kozma R., Lee J., Monfries C., Harden N., Lim L. Biochem. J. 280:233-241(1991).

30   [6] Ahmed S., Maruyama I.N., Kozma R., Lee J., Brenner S., Lim L. Biochem. J. 287:995-999(1992).

[7] Boguski M.S., Bairoch A., Attwood T.K., Michaels G.S. Nature 358:113-113(1992).

932. 3-dehydroquinate synthase (DHQ_synthase)

[1] Barten R, Meyer TF; Medline: 98273626 Cloning and characterisation of the Neisseria gonorrhoeae aroB gene." Mol Gen Genet 1998;258:34-44.

[2] Hawkins AR, Lamb HK; Medline: 96048023 The molecular biology of multidomain

5 proteins. Selected examples." Eur J Biochem 1995;232:7-18.

The 3-dehydroquinate synthase EC:4.6.1.3 domain is present in isolation in various bacterial 3-dehydroquinate synthases and also present as a domain in the pentafunctional AROM polypeptide Swiss:P07547 [2]. 3-dehydroquinate (DHQ) synthase catalyses the formation of

10 dehydroquinate (DHQ) and orthophosphate from 3-deoxy-D-arabino heptulosonic 7 phosphate [1]. This reaction is part of the shikimate pathway which is involved in the biosynthesis of aromatic amino acids.

Number of members: 25

15 933. Dihydrofolate reductase signature (DiHfolate_red)

Dihydrofolate reductases (EC 1.5.1.3) [1] are ubiquitous enzymes which catalyze the reduction of folic acid into tetrahydrofolic acid. They can be inhibited by a number of antagonists such as trimethroprim and methotrexate which are used as antibacterial or

20 anticancerous agents. A signature pattern was derived from a region in the N-terminal part of these enzymes, which includes a conserved Pro-Trp dipeptide; the tryptophan has been shown [2] to be involved in the binding of substrate by the enzyme.

Consensus pattern[LVAGC]-[LIF]-G-x(4)-[LIVMF]-P-W-x(4,5)-[DE]-x(3)-[FYIV]-

25 x(3)-[STIQ] Sequences known to belong to this class detected by the patternALL, except for type II bacterial, plasmid-encoded, dihydrofolate reductases which do not belong to the same class of enzymes.

[1] Harpers' Review of Biochemistry, Lange, Los Altos (1985).

30 [2] Bolin J.T., Filman D.J., Matthews D.A., Hamlin R.C., Kraut J. J. Biol. Chem. 257:13650-13662(1982).

934. (DIL)

[1] Ponting CP; Medline: 95397417 AF-6/cno: neither a kinesin nor a myosin, but a bit of both." Trends Biochem Sci 1995;20:265-266.

Number of members: 31

935. (DNA_gyraseB_C)

DNA topoisomerase II signature (cross-reference = TOPOISOMERASE_II)

DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type II topoisomerases are ATP-dependent and act by passing a DNA segment through a transient double-strand break. Topoisomerase II is found in phages, archaebacteria, prokaryotes, eukaryotes, and in African Swine Fever virus (ASF). In bacteriophage T4 topoisomerase II consists of three subunits (the product of genes 39, 52 and 60). In prokaryotes and in archaebacteria the enzyme, known as DNA gyrase, consists of two subunits (genes gyrA and gyrB [E2]). In some bacteria, a second type II topoisomerase has been identified; it is known as topoisomerase IV and is required for chromosome segregation, it also consists of two subunits (genes parC and parE). In eukaryotes, type II topoisomerase is a homodimer.

There are many regions of sequence homology between the different subtypes of topoisomerase II. The relation between the different subunits is shown in the following representation:

```
        <---------------About-1400-residues--------------------->
        [----------Protein 39-*-----][----Protein 52----] Phage T4
        [----------gyrB-------*-----][--------gyrA----------------] Prokaryote II
         Archaebacteria
        [----------parE-------*-----][--------parD----------------] Prokaryote IV
        [--------------------*---------------------------------] Eukaryote and ASF
```

'*': Position of the pattern.

As a signature pattern for this family of proteins, a region was selected that contains a highly conserved pentapeptide. The pattern is located in gyrB, in parE, and in protein 39 of phage T4 topoisomerase.

Consensus pattern [LIVMA]-x-E-G-[DN]-S-A-x-[STAG] Sequences known to belong to this class detected by the pattern ALL.

5 [1] Sternglanz R. Curr. Opin. Cell Biol. 1:533-535(1990).

[2] Bjornsti M.-A. Curr. Opin. Struct. Biol. 1:99-103(1991).

[3] Sharma A., Mondragon A. Curr. Opin. Struct. Biol. 5:39-47(1995).

[4] Roca J. Trends Biochem. Sci. 20:156-160(1995).

10 936. (DNA_topoisolIV)

DNA topoisomerase II signature (cross-reference = TOPOISOMERASE_II)

DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type II topoisomerases are ATP-

15 dependent and act by passing a DNA segment through a transient double-strand break. Topoisomerase II is found in phages, archaebacteria, prokaryotes, eukaryotes, and in African Swine Fever virus (ASF). In bacteriophage T4 topoisomerase II consists of three subunits (the product of genes 39, 52 and 60). In prokaryotes and in archaebacteria the enzyme, known as DNA gyrase, consists of two subunits (genes gyrA and gyrB [E2]). In some bacteria, a

20 second type II topoisomerase has been identified; it is known as topoisomerase IV and is required for chromosome segregation, it also consists of two subunits (genes parC and parE). In eukaryotes, type II topoisomerase is a homodimer.

There are many regions of sequence homology between the different subtypes of

25 topoisomerase II. The relation between the different subunits is shown in the following representation:

```
<---------------About-1400-residues--------------------->
[----------Protein 39-*-----][----Protein 52----] Phage T4
30  [----------gyrB-------*-----][-------gyrA----------------] Prokaryote II Archaebacteria
[----------parE-------*-----][-------parD----------------] Prokaryote IV
[-------------------*-----------------------------------] Eukaryote and ASF
'*': Position of the pattern.
```

As a signature pattern for this family of proteins, a region was selected that contains a highly conserved pentapeptide. The pattern is located in gyrB, in parE, and in protein 39 of phage T4 topoisomerase.

5   Consensus pattern [LIVMA]-x-E-G-[DN]-S-A-x-[STAG] Sequences known to belong to this class detected by the patternALL.

[1] Sternglanz R. Curr. Opin. Cell Biol. 1:533-535(1990).

[2] Bjornsti M.-A. Curr. Opin. Struct. Biol. 1:99-103(1991).

10   [3] Sharma A., Mondragon A. Curr. Opin. Struct. Biol. 5:39-47(1995).

[4] Roca J. Trends Biochem. Sci. 20:156-160(1995).

937. Prolyl oligopeptidase family serine active site (DPPIV_N_term)

15   The prolyl oligopeptidase family [1,2,3] consist of a number of evolutionary related peptidases whose catalytic activity seems to be provided by a charge relay system similar to that of the trypsin family of serine proteases, but which evolved by independent convergent evolution. The known members of this family are listed below.

    - Prolyl endopeptidase (EC 3.4.21.26) (PE) (also called post-proline cleaving enzyme). PE is

20   an enzyme that cleaves peptide bonds on the C-terminal side of prolyl residues. The sequence of PE has been obtained from a mammalian species (pig) and from bacteria (Flavobacterium meningosepticum and Aeromonas hydrophila); there is a high degree of sequence conservation between these sequences.

    - Escherichia coli protease II (EC 3.4.21.83) (oligopeptidase B) (gene prtB) which cleaves

25   peptide bonds on the C-terminal side of lysyl and argininyl residues.

    - Dipeptidyl peptidase IV (EC 3.4.14.5) (DPP IV). DPP IV is an enzyme that removes N-terminal dipeptides sequentially from polypeptides having unsubstituted N-termini provided that the penultimate residue is proline.

    - Yeast vacuolar dipeptidyl aminopeptidase A (DPAP A) (gene: STE13) which is responsible

30   for the proteolytic maturation of the alpha-factor precursor.

    - Yeast vacuolar dipeptidyl aminopeptidase B (DPAP B) (gene: DAP2).

    - Acylamino-acid-releasing enzyme (EC 3.4.19.1) (acyl-peptide hydrolase). This enzyme catalyzes the hydrolysis of the amino-terminal peptide bond of an N-acetylated protein to generate a N-acetylated amino acid and a protein with a free amino-terminus.

A conserved serine residue has experimentally been shown (in E.coli protease II as well as in pig and bacterial PE) to be necessary for the catalytic mechanism. This serine, which is part of the catalytic triad (Ser, His, Asp), is generally located about 150 residues away from the C-

5   terminal extremity of these enzymes (which are all proteins that contains about 700 to 800 amino acids).

Consensus pattern D-x(3)-A-x(3)-[LIVMFYW]-x(14)-G-x-S-x-G-G-[LIVMFYW](2) [S is the active site residue] Sequences known to belong to this class detected by the pattern ALL,

10   except for yeast DPAP A.

Note: these proteins belong to families S9A/S9B/S9C in the classification of peptidases [4,E1].

15   [1] Rawlings N.D., Polgar L., Barrett A.J. Biochem. J. 279:907-911(1991).

[2] Barrett A.J., Rawlings N.D. Biol. Chem. Hoppe-Seyler 373:353-360(1992).

[3] Polgar L., Szabo E. Biol. Chem. Hoppe-Seyler 373:361-366(1992).

[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

20   938. Deoxyhypusine synthase (DS)

Eukaryotic initiation factor 5A (eIF-5A) contains an unusual amino acid, hypusine [N epsilon-(4-aminobutyl-2-hydroxy)lysine]. The first step in the post-translational formation of hypusine is catalysed by the enzyme

25   deoxyhypusine synthase (DS) EC:1.1.1.249. The modified version of eIF-5A, and DS, are required for eukaryotic cell proliferation [1].

Number of members: 9

[1] Liao DI, Wolff EC, Park MH, Davies DR; Medline: 98154315 Crystal structure of the

30   NAD complex of human deoxyhypusine synthase: an enzyme with a ball-and-chain mechanism for blocking the active site." Structure 1998;6:23-32.

939. (DUF21)

Many of the sequences in this family are annotated as hemolysins, however this is due to a similarity to Swiss:Q54318 that does not contain this domain. This domain is found in the N-terminus of the proteins adjacent to two intracellular CBS domains CBS.

Number of members: 42

940. (DUF59)

This family includes prokaryotic proteins of unknown function. The family also includes PhaH Swiss:O84984 from Pseudomonas putida. PhaH forms a complex with PhaF Swiss:O84982, PhaG Swiss:O84983 and PhaI Swiss:O84985, which hydroxylates phenylacetic acid to 2-hydroxyphenylacetic acid [1]. So members of this family may all be components of ring hydroxylating complexes.

Number of members: 15

[1] Olivera ER, Minambres B, Garcia B, Muniz C, Moreno MA, Ferrandez A, Diaz E, Garcia JL, Luengo JM; Medline: 98263372   Molecular characterization of the phenylacetic acid catabolic pathway in Pseudomonas putida U: the phenylacetyl-CoA catabolon." Proc Natl Acad Sci U S A 1998;95:6419-6424.

941. (DUF82)

The protein contains four conserved cysteines that may be involved in metal binding or disulphide bridges.

Number of members: 4

942. Riboflavin kinase / FAD synthetase (FAD_Synth)

This family consists part of the bifunctional enzyme riboflavin kinase / FAD synthetase. These enzymes have both ATP:riboflavin 5'-phospho transferase and ATP:FMN-adenylyltransferase activitys [1]. They catalyse the 5'-phosphorylation of riboflavin to FMN and the adenylylation of FMN to FAD [1].

CAUTION: It is not clear if this region of the enzymes catalyses either or both of the enzymatic reactions.
Number of members: 27

[1] Manstein DJ, Pai EF; Medline: 87057286   Purification and characterization of FAD synthetase from Brevibacterium ammoniagenes." J Biol Chem 1986;261:16169-16173.

943. [2Fe-2S] binding domain (fer2_2)

[1] Romao MJ, Archer M, Moura I, Moura JJ, LeGall J, Engh R, Schneider M, Hof P, Huber R; Medline: 96072968   Crystal structure of the xanthine oxidase-related aldehyde oxido-reductase from D. gigas." Science 1995;270:1170-1176.
Number of members: 53

944. Filovirus glycoprotein (Filo_glycop)

This family includes an extracellular region from the envelope glycoprotein of Ebola and Marburg viruses. This region is also produced as a separate transcript that gives rise to a non-structural, secreted glycoprotein, which is produced in large amounts and has an unknown function [1]. Processing of this protein may be involved in viral pathogenicity [2].
Number of members: 23

[1] Volchkov VE, Feldmann H, Volchkova VA, Klenk HD; Medline: 98245155   Processing of the Ebola virus glycoprotein by the proprotein convertase furin." Proc Natl Acad Sci U S A 1998;95:5762-5767.
[2] Sanchez A, Trappier SG, Mahy BW, Peters CJ, Nichol ST; Medline: 96195018   The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing." Proc Natl Acad Sci U S A 1996;93:3602-3607.

945. Frataxin-like domain (Frataxin_Cyay)

This family contains proteins that have a domain related to the globular C-terminus of Frataxin the protein that is mutated in Friedreich's ataxia. This domain is found in a family of bacterial proteins. The function of this domain is currently unknown.

Number of members: 12

[1] Gibson TJ, Koonin EV, Musco G, Pastore A, Bork P; Medline: 97084946   Friedreich's
ataxia protein: phylogenetic evidence for mitochondrial dysfunction." Trends Neurosci
1996;19:465-468.

946. (GAF)

Domain present in phytochromes and cGMP-specific phosphodiesterases.

Number of members: 296

[1] Aravind L, Ponting CP; Medline: 98094688   The GAF domain: an evolutionary link
between diverse phototransducing proteins." Trends Biochem Sci 1997;22:458-459.

947. Galaptin signature (Gal-bind_lectin)

All vertebrates synthesize soluble galactoside-binding lectins [1,2,3] (also known as
galectins, galaptins or S-lectin). These carbohydrate-binding proteins are developmentally
regulated. Although their exact physiological role is not yet clear they seem to be involved in
differentiation, cellular regulation and tissue construction. The sequence of galactoside-
binding lectins from electric eel (electrolectin), conger eel (congerin), chicken and a number
of mammalian species is known. These lectins are proteins of about 130 to 140 amino acid
residues (14 Kd to 16 Kd).

A number of other proteins are known to belong to this family:
   - Galectin-3 (also known as MAC-2 antigen; CBP-35 or IgE-binding protein), a 35 Kd lectin
which binds immunoglobulin E and which is composed of two domains: a N-terminal domain
that consist of tandem repeats of a glycine/ proline-rich sequence and a C-terminal galaptin
domain.
   - Galectin-4 [4], which is composed of two galaptin domains.
   - Galectin-5.
   - Galectin-7 [5], a keratinocyte protein which could be involved in cell-cell and/or cell-
matrix interactions necessary for normal growth control.
   - Galectin-8 [6], which is composed of two galaptin domains.

- Galectin-9 [7], which is composed of two galaptin domains.

- Human eosinophil lysophospholipase (EC 3.1.1.5) [8] (Charcot-Leyden crystal protein), a protein that may have both an enzymatic and a lectin activities. It forms hexagonal bipyramidal crystals in tissues and secretions from sites of eosinophil-associated

5 inflammation.

- Caenorhabditis elegans 32 Kd lactose-binding lectin [9]. This lectin is composed of two galaptin domains.

- Caenorhabditis elegans lec-7 and lec-8.

One of the conserved regions of these lectins contains a tryptophan that has been shown [10]

10 to be essential to the binding of galactosides. This region was used as a signature pattern for these proteins.

Consensus patternW-[GEK]-x-[EQ]-x-[KRE]-x(3,6)-[PCTF]-[LIVMF]-[NQEGSKV]-x-[GH]-x(3)-[DENKHS]-[LIVMFC] [W binds carbohydrate] Sequences known to belong to

15 this class detected by the pattern ALL, except for pig galectin 4.

[1] Barondes S.H., Gitt M.A., Leffler H., Cooper D.N.W. Biochimie 70:1627-1632(1988).

[2] Hirabayashi J., Kasai K.-I. J. Biochem. 104:1-4(1988).

[3] Barondes S.H., Castronovo V., Cooper D.N.W., Cummings R.D., Drickamer K., Feizi

20 T., Gitt M.A., Hirabayashi J., Hughes C., Kasai K.-I., Leffler H., Liu F.-T., Lotan R., Mercurio A.M., Monsigny M., Pillair S., Poirer F., Raz A., Rigby P.W.J., Rini J.M., Wang J.L. Cell 76:597-598(1994).

[4] Oda Y., Herrmann J., Gitt M., Turck C.W., Burlingame A.L., Barondes S.H., Leffler H. J. Biol. Chem. 268:5929-5939(1993).

25 [5] Madsen P., Rasmussen H.H., Flint T., Gromov P., Kruse T.A., Honore B., Vorum H., Celis J.E. J. Biol. Chem. 270:5823-5829(1995).

[6] Hadari Y.R., Paz K., Dekel R., Mestrovic T., Accili D., Zick Y. J. Biol. Chem. 270:3447-3453(1995).

[7] Wada J., Kanwar Y.S. J. Biol. Chem. 272:6078-6086(1997).

30 [8] Ackerman S.J., Corrette S.E., Rosenberg H.F., Bennett J.C., Mastrianni D.M., Nicholson-Weller A., Weller P.F., Chin D.T., Tenen D.G. J. Immunol. 150:456-468(1993).

[9] Hirabayashi J., Satoh M., Kasai K.-I. J. Biol. Chem. 267:15485-15490(1992).

[10] Abbott W.M., Feizi T. J. Biol. Chem. 266:5552-5557(1991).

948. (GARS) Phosphoribosylglycinamide synthetase signature (phosphoribosylamine glycine ligase)

PROSITE: PDOC00164; cross-reference(s): PS00184

5   [1] catalyzes the second step in the de novo biosynthesis of purine, the ATP-dependent addition of 5-phosphoribosylamine to glycine to form 5'phosphoribosylglycinamide.

In bacteria GARS is a monofunctional enzyme (encoded by the purD gene), in of a bifunctional enzyme (encoded by the ADE5,7 gene), in higher eukaryotes it is part, with AIRS and with phosphoribosylglycinamide formyltransferase (GART) of a trifunctional

10 enzyme (GARS-AIRS-GART).

The sequence of GARS is well conserved. A highly conserved octapeptide was selected as a signature pattern.

Consensus patternR-F-G-D-P-E-x-[QM]

15 Sequences known to belong to this class detected by the patternALL.

[1]Aiba A., Mizobuchi K. J. Biol. Chem. 264:21239-21246(1989).

949. GLTT - GLTT repeat (12 copies)

20 This short repeat of unknown function is found in multiple copies in several C. elegans proteins. The repeat is five residues long and consists of XGLTT where X can be any amino acid. Number of members: 34.

950. Glu_synthase - Conserved region in glutamate synthase

25 This family represents a region of the glutamate synthase protein. This region is expressed as a seperate subunit in the glutamate synthase alpha subunit from archaebacteria, or part of a large multidomain enzyme in other organisms. The aligned region of these proteins contains a putative FMN binding site and Fe-S cluster. Number of members: 44.

30 [1] Medline: 97082505. Sequence of the GLT1 gene from Saccharomyces cerevisiae reveals the domain structure of yeast glutamate synthase. Filetici P, Martegani MP, Valenzuela L, Gonzalez A, Ballario P; Yeast 1996;12:1359-1366.

951. (Glyco_hydro_2) Glycosyl hydrolases family 2 signatures

GLYCOSYL_HYDROL_F2_1; PS00608; GLYCOSYL_HYDROL_F2_2

It has been shown [1,2,E1] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:

5 -Beta-galactosidases (EC 3.2.1.23) from bacteria such as Escherichia coli (genes lacZ and ebgA), Clostridium acetobutylicum, Clostridium thermosulfurogenes, Klebsiella pneumoniae, Lactobacillus delbrueckii, or Streptococcus thermophilus and from the fungi Kluyveromyces lactis.

-Beta-glucuronidase (EC 3.2.1.31) from Escherichia coli (gene uidA) and from mammals. One of the conserved regions in these enzymes is centered on a conserved glutamic acid

10 residue which has been shown [3], in Escherichia coli lacZ, to be the general acid/base catalyst in the active site of the enzyme. This region has been used as a signature pattern. A highly conserved region located some sixty residues upstream from the active site glutamate has been selected as a second signature pattern.

15 Consensus pattern N-x-[LIVMFYWD]-R-[STACN](2)-H-Y-P-x(4)-[LIVMFYWS](2)-x(3)-[DN]-x(2)-G-[LIVMFYW](4) Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern [DENQLF]-[KRVW]-N-[HRY]-[STAPPV]-[SAC]-[LIVMFS](3)-W-
20 [GS]-x(2,3)-N-E [E is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for Rhizobium meliloti lacZ.

[1]Henrissat B. Biochem. J. 280:309-316(1991).

[2]Schroeder C.J., Robert C., Lenzen G., McKay L.L., Mercenier A. J. Gen. Microbiol.
25 137:369-380(1991).

[3]Gebler J.C., Aebersold R., Withers S.G. J. Biol. Chem. 267:11126-11130(1992).

952. (Glyco_hydro_3) Glycosyl hydrolases family 3 active site

PROSITE: PDOC00621. PROSITE cross-reference(s)PS00775; GLYCOSYL_HYDROL_F3

30   It has been shown [1,2] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:

-Beta glucosidases (EC 3.2.1.21) from the fungi Aspergillus wentii (A-3), Hansenula anomala, Kluyveromyces fragilis, Saccharomycopsis fibuligera,(BGL1 and BGL2), Schizophyllum commune and Trichoderma reesei (BGL1).

-Beta glucosidases from the bacteria Agrobacterium tumefaciens (Cbg1), Butyrivibrio fibrisolvens (bglA), Clostridium thermocellum (bglB), Escherichia coli (bglX), Erwinia chrysanthemi (bgxA) and Ruminococcus albus.

-Alteromonas strain O-7 beta-hexosaminidase A (EC 3.2.1.52).

5   -Bacillus subtilis hypothetical protein yzbA.

-Escherichica coli hypothetical protein ycfO and HI0959, the corresponding Haemophilus influenzae protein.

One of the conserved regions in these enzymes is centered on a conserved aspartic acid residue which has been shown [3], in Aspergillus wentii beta-glucosidase A3, to be

10   implicated in the catalytic mechanism. This region was used as a signature pattern.

Consensus pattern[LIVM](2)-[KR]-x-[EQK]-x(4)-G-[LIVMFT]-[LIVT]-[LIVMF]-[ST]-D-x(2)-[SGADNI] [D is the active site residue]

Sequences known to belong to this class detected by the patternALL.

15

[1]Henrissat B. Biochem. J. 280:309-316(1991).

[2]Castle L.A., Smith K.D., Morris R.O. J. Bacteriol. 174:1478-1486(1992).

[3]Bause E., Legler G. Biochim. Biophys. Acta 626:459-465(1980).

20   953. GP120 - Envelope glycoprotein GP120

The entry of HIV requires interaction of viral GP120 with Swiss:P01730 and a chemokine receptor on the cell surface. Number of members: 17891

[1]Medline: 98303379. Structure of an HIV gp120 envelope glycoprotein in complex with

25   the CD4 receptor and a neutralizing human antibody. Kwong PD, Wyatt R, Robinson J, Sweet RW, Sodroski J, Hendrickson WA; Nature 1998;393:648-659.

954. (GSPII_E) Bacterial type II secretion system protein E signature

PROSITE: PDOC00567. PROSITE cross-reference(s) PS00662; T2SP_E

30   A number of bacterial proteins, some of which are involved in a general secretion pathway (GSP) for the export of proteins (also called the type II pathway) [1,2], have been found to be evolutionary related. These proteins are listed below:

-The 'E' protein from the GSP operon of: Aeromonas (gene exeE); Erwinia (gene outE); Escherichia coli (gene yheG); Klebsiella pneumoniae (gene pulE); Pseudomonas aeruginosa (gene xcpR); Vibrio cholerae (gene epsE) and Xanthomonas campestris (gene xpsE).

-Agrobacterium tumefaciens Ti plasmid virB operon protein 11. This protein is required for

5   the transfer of T-DNA to plants.

-Bacillus subtilis comG operon protein 1 which is required for the uptake of DNA by competent Bacillus subtilis cells.

-Aeromonas hydrophila tapB, involved in type IV pilus assembly.

-Pseudomonas protein pilB, which is essential for the formation of the pili.

10   -Pseudomonas aeruginosa protein twitching mobility protein pilT.

-Neisseria gonorrhoeae type IV pilus assembly protein pilF.

-Vibrio cholerae protein tcpT, which is involved in the biosynthesis of the tcp pilus.

-Escherichia coli protein hofB (hopB).

15   -Escherichia coli hypothetical protein ygcB.

-Escherichia coli hypothetical protein yggR.

These proteins have from 344 (pilT and virB11) to 568 (tapB) amino acids, they are probably cytoplasmically located and, on the basis of the presence of a conserved P-loop region (see <PDOC00017>), probably bind ATP. A region that overlaps the 'B' motif of

20   ATP-binding proteins was selected as a signature pattern.

Consensus pattern[LIVM]-R-x(2)-P-D-x-[LIVM](3)-G-E-[LIVM]-R-D

Sequences known to belong to this class detected by the patternALL, except for ygcB.

25   [1]Salmond G.P.C., Reeves P.J. Trends Biochem. Sci. 18:7-12(1993).

[2]Hobbs M., Mattick J.S. Mol. Microbiol. 10:233-243(1993).

955. (guanylate_cyc) Guanylate cyclases signature

PROSITE: PDOC00425. PROSITE cross-reference(s) PS00452;

30   GUANYLATE_CYCLASES   Guanylate cyclases (EC 4.6.1.2) [1 to 4] catalyze the formation of cyclic GMP (cGMP) from GTP. cGMP acts as an intracellular messenger, activating cGMP dependent kinases and regulating CGMP-sensitive ion channels. The role of cGMP as a second messenger in vascular smooth muscle relaxation and retinal photo-transduction is well established. Guanylate cyclase is found both in the soluble and particular

fraction of eukaryotic cells. The soluble and plasma membrane-bound forms differ in structure, regulation and other properties.

Most currently known plasma membrane-bound forms are receptors for small polypeptides. The topology of such proteins is the following: they have a N-terminal
5   extracellular domain which acts as the ligand binding region, then a transmembrane domain, followed by a large cytoplasmic C-terminal region that can be subdivided into two domains: a protein kinase-like domain that appears important for proper signalling and a cyclase catalytic domain. This topology is schematically represented below.

```
10    +-------------------xxxxx--------------------+--------------+
      | Ligand-binding   XXXXX Protein Kinase like | Cyclase  |
      +-------------------xxxxx--------------------+--------------+
      Extracellular   Transmembrane   Cytoplasmic
```

15   The known guanylate cyclase receptors are:

-The sea-urchins receptors for speract and resact, which are small peptides that stimulate sperm motility and metabolism.

-The receptors for natriuretic peptides (ANF). Two forms of ANF receptors with guanylate cyclase activity are currently known: GC-A (or ANP-A) which seems specific to atrial
20   natriuretic peptide (ANP), and GC-B (or ANP-B) which seems to be stimulated more effectively by brain natriuretic peptide (BNP) than by ANP.

-The receptor for Escherichia coli heat-stable enterotoxin (GC-C). The endogenous ligand for this intestinal receptor seems to be a small peptide called guanylin.

-Retinal guanylate cyclase (retGC) which probably plays a specific functional role in the
25   rods and/or cones of photoreceptors. It is not known if this protein acts as receptor, but its structure is similar to that of the other plasma membrane-bound GCs.

The soluble forms of guanylate cyclase are cytoplasmic heterodimers. The two subunits, alpha and beta are proteins of from 70 to 82 Kd which are highly related. Two forms of beta subunits are currently known: beta-1 which seems to be expressed in lung and
30   brain, and beta-2 which is more abundant in kidney and liver.

The membrane and cytoplasmic forms of guanylate cyclase share a conserved domain which is probably important for the catalytic activity of the enzyme. Such a domain is also found twice in the different forms of membrane-bound adenylate cyclases (also known as

class-III) [5,6] from mammals, slime mold or Drosophila. A consensus pattern was derived from the most conserved region in that domain.

Consensus patternG-V-[LIVM]-x(0,1)-G-x(5)-[FY]-x-[LIVM]-[FYW]-[GS]-[DNTHKW]-
5   [DNT]-[IV]-[DNTA]-x(5)-[DE]

Sequences known to belong to this class detected by the patternALL, except for the sea urchin Arbacia punctulata resact receptor which lack this domain.

Note this pattern will detect both domains of adenylate cyclases class-III.

10   [1]Koesling D., Boehme E., Schultz G. FASEB J. 5:2785-2791(1991).

[2]Garbers D.L. New Biol. 2:499-504(1990).

[3]Garbers D.L. Cell 71:1-4(1992).

[4]Yuen P.S.T., Garbers D.L. Annu. Rev. Neurosci. 15:193-225(1992).

[5]Iyengar R. FASEB J. 7:768-775(1993).

15   [6]Barzu O., Danchin A. Prog. Nucleic Acid Res. Mol. Biol. 49:241-283(1994).

956. Hemolysin-type calcium-binding region signature (HemolysinCabinD)

Gram-negative bacteria produce a number of proteins which are secreted into the growth
20   medium by a mechanism that does not require a cleaved N-terminal signal sequence. These proteins, while having different functions, seem [1] to share two properties: they bind calcium and they contain a variable number of tandem repeats consisting of a nine amino acid motif rich in glycine, aspartic acid and asparagine. It has been shown [2] that such a domain is involved in the binding of calcium ions in a parallel beta roll structure. The proteins which
25   are currently known to belong to this category are:

- Hemolysins from various species of bacteria. Bacterial hemolysins are exotoxins that attack blood cell membranes and cause cell rupture. The hemolysins which are known to contain such a domain are those from: E. coli (gene hlyA), A. pleuropneumoniae (gene appA), A. actinomycetemcomitans and P. haemolytica (leukotoxin) (gene lktA).

30   - Cyclolysin from Bordetella pertussis (gene cyaA). A multifunctional protein which is both an adenylate cyclase and a hemolysin.

- Extracellular zinc proteases: serralysin (EC 3.4.24.40) from Serratia, prtB and prtC from Erwinia chrysanthemi and aprA from Pseudomonas aeruginosa.

- Nodulation protein nodO from Rhizobium leguminosarum.

A signature pattern was derived from conserved positions in the sequence of the calcium-binding domain.

Consensus pattern D-x-[LI]-x(4)-G-x-D-x-[LI]-x-G-G-x(3)-D Sequences known to belong to

5 this class detected by the pattern ALL.

Note: This pattern is found once in nodO and the extracellular proteases but up to 5 times in some hemolysin/cyclolysins.

10 [1] Economou A., Hamilton W.D.O., Johnston A.W.B., Downie J.A. EMBO J. 9:349-354(1990).

[2] Baumann U., Wu S., Flaherty K.M., McKay D.B. EMBO J. 12:3357-3364(1993).

957. Hint module (Hint)

15

This is an alignment of the Hint module in the Hedgehog proteins. It does not include any Inteins which also possess the Hint module.

Number of members: 36

20 [1] Hall TM, Porter JA, Young KE, Koonin EV, Beachy PA, Leahy DJ; Medline: 97474313

Crystal structure of a Hedgehog autoprocessing domain: homology between Hedgehog and self-splicing proteins." Cell 1997;91:85-97.

958. Hydantoinase/oxoprolinase (Hydantoinase)

25

This family includes the enzymes hydantoinase and oxoprolinase EC:3.5.2.9. Both reactions involve the hydrolysis of 5-membered rings via hydrolysis of their internal imide bonds [1].

Number of members: 14

30 [1] Ye GJ, Breslow EB, Meister A, Guo-jie GE$[corrected to Ye GJ]; Medline: 97113037

The amino acid sequence of rat kidney 5-oxo-L-prolinase determined by cDNA cloning" [published erratum appears in J Biol Chem 1997 Feb 14;272(7):4646] J Biol Chem 1996;271:32293-32300.

959. IMP dehydrogenase / GMP reductase signature (IMPDH_N)

IMP dehydrogenase (EC 1.1.1.205) (IMPDH) catalyzes the rate-limiting reaction of de novo

GTP biosynthesis, the NAD-dependent reduction of IMP into XMP [1]. Inhibition of IMP

5   dehydrogenase activity results in the cessation of DNA synthesis. As IMP dehydrogenase is

associated with cell proliferation, it is a possible target for cancer chemotherapy. Mammalian

and bacterial IMPDHs are tetramers of identical chains. There are two IMP dehydrogenase

isozymes in humans [2].

10   GMP reductase (EC 1.6.6.8) catalyzes the irreversible and NADPH-dependent reductive

deamination of GMP into IMP [3]. It converts nucleobase, nucleoside and nucleotide

derivatives of G to A nucleotides, and maintains intracellular balance of A and G nucleotides.

IMP dehydrogenase and GMP reductase share many regions of sequence similarity. One of

15   these regions is centered on a cysteine residue thought [3] to be involved in binding IMP.

This region was used as a signature pattern.

Consensus pattern[LIVM]-[RK]-[LIVM]-G-[LIVM]-G-x-G-S-[LIVM]-C-x-T [C is the

putative IMP-binding residue] Sequences known to belong to this class detected by the

20   pattern ALL.

[1] Collart F.R., Huberman E. J. Biol. Chem. 263:15769-15772(1988).

[2] Natsumeda Y., Ohno S., Kawasaki H., Konno Y., Weber G., Suzuki K. J. Biol. Chem.

265:5292-5295(1990).

25   [3] Andrews S.C., Guest J.R. Biochem. J. 255:35-43(1988).

960. impB/mucB/samB family (IMS)

These proteins are involved in UV protection (Swiss).

30   Number of members: 38

961. Type II intron maturase (Intron_maturas2)

Group II introns use intron-encoded reverse transcriptase, maturase and DNA endonuclease activities for site-specific insertion into DNA [2]. Although this type of intron is self splicing in vitro they require a maturase protein for

splicing in vivo. It has been shown that a specific region of the al2 intron is needed for the

5   maturase function [1]. This region was found to be conserved in group II introns and called domain X [3].

Number of members: 335

[1] Moran JV, Mecklenburg KL, Sass P, Belcher SM, Mahnke D, Lewin A, Perlman P;

10   Medline: 94301788   Splicing defective mutants of the COXI gene of yeast mitochondrial DNA: initial definition of the maturase domain of the group II intron al2. Nucleic Acids Res 1994;22:2057-2064.

[2] Guo H, Zimmerly S, Perlman PS, Lambowitz AM; Medline: 98031910   Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in

15   double-stranded DNA." EMBO J 1997;16:6835-6848.

[3] Mohr G, Perlman PS, Lambowitz AM; Medline: 94077696   Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function." Nucleic Acids Res 1993;21:4991-4997.

20   962. LAGLIDADG endonuclease (Intron_maturase)

[1] Heath PJ, Stephens KM, Monnat RJ Jr, Stoddard BL; Medline: 97331323   The structure of I-CreI, a group I intron-encoded homing endonuclease." Nat Struct Biol 1997;4:468-476.

[2] Belfort M, Roberts RJ; Medline: 97402526   Homing endonucleases: keeping the house in

25   order." Nucleic Acids Res 1997;25:3379-3388.

[3] Dalgaard JZ, Klar AJ, Moser MJ, Holley WR, Chatterjee A, Mian IS; Medline: 98026854 Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family." Nucleic Acids Res 1997;25:4626-4638.

30

Number of members: 220

963. Isopentenyl transferase (IPT)

Isopentenyl transferase / dimethylallyl transferase synthesizes isopentenyladensosine 5'-monophosphate, a cytokinin that induces shoot formation on host plants infected with the Ti plasmid [1].

Number of members: 16

[1] Canaday J, Gerad JC, Crouzet P, Otten L; Medline: 93101133 "Organization and functional analysis of three T-DNAs from the vitopine Ti plasmid pTiS4." Mol Gen Genet 1992;235:292-303.

964. Laminin EGF-like (Domains III and V) (laminin_EGF)

This family is like EGF but has 8 conserved cysteines instead of 6.

Number of members: 501

[1] Engel J; Medline: 93041759   Laminins and other strange proteins." Biochemistry 1992;31:10643-10651.

965. Legume lectins signatures (lectin_legA)

Leguminous plants synthesize sugar-binding proteins which are called legume lectins [1,2]. These lectins are generally found in the seeds. The exact function of legume lectins is not known but they may be involved in the attachment of nitrogen-fixing bacteria to legumes and in the protection against pathogens. Legume lectins bind calcium and manganese (or other transition metals).

Legume lectins are synthesized as precursor proteins of about 230 to 260 amino acid residues. Some legume lectins are proteolytically processed to produce two chains: beta (which corresponds to the N-terminal) and alpha (C-terminal). The lectin concanavalin A (conA) from jack bean is exceptional in that the two chains are transposed and ligated (by formation of a new peptide bond). The N-terminus of mature conA thus corresponds to that of the alpha chain and the C-terminus to the beta chain.

Two signature patterns were developed specific to legume lectins: the first is located in the C-terminal section of the beta chain and contains a conserved aspartic acid residue important for

the binding of calcium and manganese; the second one is located in the N-terminal of the alpha chain.

Consensus pattern [LIV]-[STAG]-V-[DEQV]-[FLI]-D-[ST] [D binds manganese and

5   calcium] Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern [LIV]-x-[EDQ]-[FYWKR]-V-x-[LIVF]-G-[LF]-[ST] Sequences known to belong to this class detected by the pattern ALL.

10   [1] Sharon N., Lis H. FASEB J. 4:3198-320(1990).

     [2] Lis H., Sharon N. Annu. Rev. Biochem. 55:33-37(1986).

966. Malate synthase signature (malate_synthase)

15   Malate synthase (EC 4.1.3.2) catalyzes the aldol condensation of glyoxylate with acetyl-CoA to form malate - the second step of the glyoxylate bypass, an alternative to the tricarboxylic acid cycle in bacteria, fungi and plants. Malate synthase is a protein of 530 to 570 amino acids whose sequence is highly conserved across species [1]. As a signature pattern, a very conserved region was selected in the central section of the enzyme.

20

Consensus pattern[KR]-[DENQ]-H-x(2)-G-L-N-x-G-x-W-D-Y-[LIVM]-F Sequences known to belong to this class detected by the pattern ALL.

    [1] Bruinenberg P.G., Blaauw M., Kazemier B., Ab G. Yeast 6:245-254(1990).

25

967. MatK/TrnK amino terminal region (MatK_N)

[1] Mohr G, Perlman PS, Lambowitz AM; Medline: 94077696 Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be

30   related to maturase function." Nucleic Acids Res 1993;21:4991-4997.

Number of members: 495

.

968. MOZ/SAS family (MOZ_SAS)

This region of these proteins has been suggested to be homologous to acetyltransferases [1]. However the similarity is not supported by standard sequence analysis.

Number of members: 15

[1] Kamine J, Elangovan B, Subramanian T, Coleman D, Chinnadurai G; Medline: 96182937 Identification of a cellular protein that specifically interacts with the essential cysteine region of the HIV-1 Tat transactivator." Virology 1996;216:357-366.

[2] Reifsnyder C, Lowell J, Clarke A, Pillus L; Medline: 96376969 Yeast SAS silencing genes and human genes associated with AML and HIV-1 Tat interactions are homologous with acetyltransferases" [see comments] [published erratum appears in Nat Genet 1997 May;16(1):109] Nat Genet 1996;14:42-49.

969. mRNA capping enzyme (mRNA_cap_enzyme)

[1] Hakansson K, Doherty AJ, Shuman S, Wigley DB; Medline: 97304383 X-ray crystallography reveals a large conformational change during guanyl transfer by mRNA capping enzymes." Cell 1997;89:545-553.

Number of members: 7

970. DNA mismatch repair proteins mutS family signature (MutS_C)

Mismatch repair contributes to the overall fidelity of DNA replication [1]. It involves the correction of mismatched base pairs that have been missed by the proofreading element of the DNA polymerase complex. The sequence of some proteins involved in mismatch repair in different organisms have been found to be evolutionary related [2,3]. One of these families is called mutS [4,E1], it consists of:

- Prokaroytic protein mutS protein (also called hexA in Streptococcus pneumoniae). Muts is thought to carry out the mismatch recognition step of DNA repair.

- Eukaryotic MSH1, which is involved in mitochondrial DNA repair.

- Eukaryotic MSH2, which is involved in nuclear postreplication mismatch repair. MSH2 heterodimerizes with MSH6. In man, MSH2 is involved in a form of familial hereditary nonpolyposis colon cancer (HNPCC).

759

- Eukaryotic MSH3, which is probably involved in the repair of large loops.

- Eukaryotic MSH4, which is involved in meiotic recombination.

- Eukaryotic MSH5, which is involved in meiotic recombination.

- Eukaryotic MSH6 (also known as G/T mismatch binding protein), a DNA-repair protein

5   that binds to G/T mismatches through heterodimerization with MSH2.

- Prokaryotic protein mutS2 whose function is not yet known.

- A coral (Sarcophyton glaucum) mitochondrial encoded mutS-like protein.

As a signature pattern for this class of mismatch repair proteins a region rich in glycine and

negatively charged residues was selected This region is found

10   in the C-terminal section of these proteins; about 80 residues to the C-terminal of an ATP-

binding site motif 'A' (P-loop) (see <PDOC00017>).

Consensus pattern[ST]-[LIVMF]-x-[LIVM]-x-D-E-[LIVMFY]-[GC]-[RKH]-G-[GST]- x(4)-

G Sequences known to belong to this class detected by the pattern ALL, except for mutS2.

15

[1] Modrich P. Annu. Rev. Biochem. 56:435-466(1987).

[2] Haber L.T., Walker G.C. EMBO J. 10:2707-2715(1991).

[3] New L., Liu K., Crouse G.F. Mol. Gen. Genet. 239:97-108(1993).

[4] Eisen J.A. Nucleic Acids Res. 26:4291-4300(1998).

20

971. MutS family, N-terminal putative DNA binding domain (MutS_N)

This family consists of the N-terminal region of proteins in the mutS family of DNA

mismatch repair proteins and is found associated with MutS_C located in the C-terminal

25   region. The mutS family of proteins is named after the salmonella typhimurium MutS protein

involved in mismatch repair; other members of the family included the eukaryotic MSH

1,2,3,4,5 and 6 proteins. These have various roles in DNA repair and recombination. Human

MSH has been implicated in non-polyposis colorectal carcinoma (HNPCC) and is a

mismatch binding protein [2]. The aligned region corresponds in part with domains A1, A2

30   (which may bind DNA) and B (which binds dsDNA in vitro) from T. thermophilus MutS as

characterised in [1].

Number of members: 43

972. Domain in Myosin and Kinesin Tails (MyTH4)

Domain present twice in myosin-VIIa, and also present in 3 other myosins.

[1] Chen ZY, Hasson T, Kelley PM, Schwender BJ, Schwartz MF, Ramakrishnan M,

5 Kimberling WJ, Mooseker MS, Corey DP; Medline: 97038686   Molecular cloning and

domain structure of human myosin-VIIa, the gene product defective in Usher syndrome 1B."

Genomics 1996;36:440-448.

Number of members: 21

10

973. Sodium and potassium ATPases beta subunits signatures (Na_K-ATPase)

The sodium pump (Na+,K+ ATPase), located in the plasma membrane of all animal cells [1],

is an heterotrimer of a catalytic subunit (alpha chain), a glycoprotein subunit of about 34 Kd

15 (beta chain) and a small hydrophobic protein of about 6 Kd. The beta subunit seems [2] to

regulate, through the assembly of alpha/beta heterodimers, the number of sodium pumps

transported to the plasma membrane.

Structurally the beta subunit is composed of a charged cytoplasmic domain of about 35

20 residues, followed by a transmembrane region, and a large extracellular domain that contains

three disulfide bonds and glycosylation sites. This structure is schematically represented in

the figure below.

```
+----+ +--+ +-----------+||||||
xxxxxxxxxxxxxxxxxxxxxxxxxxCxxxxCxCxxCxxxxxxxCxxxxxxxxxxxCxxxx
```

25 **** **** <-Cyt-><TM><------------Extracellular-------------------->

'C': conserved cysteine involved in a disulfide bond.

'*': position of the patterns.

30 Two isoforms of the beta subunit (beta-1 and beta-2) are currently known; they share about

50% sequence identity. Gastric (K+, H+) ATPase (proton pump) responsible for acid

production in the stomach consist of two subunits [3]; the beta chain is highly similar to the

sodium pump beta subunits. Two signature patterns were developed for beta subunits. The

first is located in the cytoplasmic domain, while the second is found in the extracellular domain and contains two of the cysteines involved in disulfide bonds.

Consensus pattern [FYW]-x(2)-[FYW]-x-[FYW]-[DN]-x(6)-[LIVM]-G-R-T-x(3)-W

5   Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern [RK]-x(2)-C-[RKQWI]-x(5)-L-x(2)-C-[SA]-G [The two C's are involved in disulfide bonds] Sequences known to belong to this class detected by the patternALL, except for the beta subunit of the sodium pump of brine shrimp whose sequence is highly

10   divergent in that region.

[1] Horisberger J.D., Lemas V., Krahenbul J.P., Rossier B.C. Annu. Rev. Physiol. 53:565-584(1991).

[2] McDonough A.A., Gerring K., Farley R.A. FASEB J. 4:1598-1605(1990).

15   [3] Toh B.-H., Gleeson P.A., Simpson R.J., Moritz R.L., Callaghan J.M., Goldkorn I., Jones C.M., Martinelli T.M., Mu F.-T., Humphris D.C., Pettitt J.M., Mori Y., Masuda T., Sobieszczuk P., Weinstock J., Mantamadiotis T., Baldwin G.S. Proc. Natl. Acad. Sci. U.S.A. 87:6418-6422(1990).

20   974. Respiratory-chain NADH dehydrogenase subunit 1 signatures (NADHdh)

Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria

25   (as a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex there are fifteen which are located in the membrane part, seven of which are encoded by the mitochondrial and chloroplast genomes of most species. The most conserved of these organelle-encoded subunits is known as subunit 1 (gene ND1 in mitochondrion, and NDH1 in chloroplast) and seems to contain the ubiquinone binding site.

30

The ND1 subunit is highly similar to subunit 4 of Escherichia coli formate hydrogenlyase (gene hycD), subunit C of hydrogenase-4 (gene hyfC). Paracoccus denitrificans NQO8 and Escherichia coli nuoH NADH-ubiquinone oxidoreductase subunits also belong to this family [3]. Two signature patterns were developed based on conserved regions of this subunit.

Consensus pattern G-[LIVMFYKRS]-[LIVMAGP]-Q-x-[LIVMFY]-x-D-[AGIM]-[LIVMFTA]- K-[LVMYST]-[LIVMFYG]-x-[KR]-[EQG] Sequences known to belong to this class detected by the patternALL, except for watermelon and Leishmania ND1.

5

Consensus pattern P-F-D-[LIVMFYQ]-[STAGPVM]-E-[GAC]-E-x-[EQ]-[LIVMS]-x(2)-G Sequences known to belong to this class detected by the pattern ALL, except for Chlamydomonas reinhardtii and Pisaster ochraceus ND1, and tobacco NDH1.

10   [1] Ragan C.I. Curr. Top. Bioenerg. 15:1-36(1987).

[2] Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).

[3] Weidner U., Geier S., Ptock A., Friedrich T., Leif H., Weiss H. J. Mol. Biol. 233:109-122(1993).

15   975. Nickel-dependent hydrogenases large subunit signatures (NiFeSe_Hases)

Hydrogenases are enzymes that catalyze the reversible activation of hydrogen and which occur widely in prokaryotes as well as in some eukaryotes. There are various types of hydrogenases, but all of them seem to contain at least one iron-sulfur cluster. They can be
20   broadly divided into two groups: hydrogenases containing nickel and, in some cases, also selenium (the [NiFe] and [NiFeSe] hydrogenases) and those lacking nickel (the [Fe] hydrogenases).

The [NiFe] and [NiFeSe] hydrogenases are heterodimer that consist of a small subunit that
25   contains a signal peptide and a large subunit. All the known large subunits seem to be evolutionary related [1]; they contain two Cys-x-x- Cys motifs; one at their N-terminal end; the other at their C-terminal end. These four cysteines are involved in the binding of nickel [2]. In the [NiFeSe] hydrogenases the first cysteine of the C-terminal motif is a selenocysteine which has experimentally been shown to be a nickel ligand [3]. Two patterns
30   were developed which are centered on the Cys-x-x-Cys motifs.

Alcaligenes eutrophus possess a NAD-reducing cytoplasmic hydrogenase (hoxS) [4]; this enzyme is composed of four subunits. Two of these subunits (beta and delta) are responsible for the hydrogenase reaction and are evolutionary related to the large and small subunits of

membrane-bound hydrogenases. The alpha subunit of coenzyme F420 hydrogenase (EC 1.12.99.1) (FRH) from archaebacterial methanogens also belongs to this family.

Consensus pattern R-G-[LIVMF]-E-x(15)-[QESM]-R-x-C-G-[LIVM]-C [The two C's are nickel ligands] Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern [FY]-D-P-C-[LIM]-[ASG]-C-x(2,3)-H [The two C's are nickel ligands] Sequences known to belong to this class detected by the pattern ALL.

[1] Menon N.K., Robbins J., Peck H.D. Jr., Chatelus C.Y., Choi E.-S., Przybyla A.E. J. Bacteriol. 172:1969-1977(1990).

[2] Volbeda A., Charon M.-H., Piras C., Hatchikian E.C., Frey M., Fontecilla-Camps J.C. Nature 373:580-587(1995).

[3] Eidsness M.K., Scott R.A., Prickrill B., der Vartaninan D.V., LeGall J., Moura I., Moura J.J.G., Peck H.D. Jr. Proc. Natl. Acad. Sci. U.S.A. 86:147-151(1989).

[4] Tran-Betcke A., Warnecke U., Boecker C., Zaborosch C., Friedrich B. J. Bacteriol. 172:2920-2929(1990).

976. NADH-Ubiquinone oxidoreductase (complex I), chain 5 C-terminus (oxidored_q1_C)

This sub-family represents a carboxyl terminal extension of oxidored_q1. Only NADH-Ubiquinone chain 5 from chloroplasts are in this family. This sub-family is part of complex I which catalyses the transfer of two electrons from NADH to ubiquinone in a reaction that is associated with proton translocation across the membrane.

Number of members: 572

[1] Walker JE; Medline: 93110040 The NADH:ubiquinone oxidoreductase (complex I) of respiratory chains." Q Rev Biophys 1992;25:253-324.

977. NADH-Ubiquinone oxidoreductase (complex I), chain 5 N-terminus (oxidored_q1_N)

This sub-family represents an amino terminal extension of oxidored_q1. Only NADH-Ubiquinone chain 5 and eubacterial chain L are in this family. This sub-family is part of

764

complex I which catalyses the transfer of two electrons from NADH to ubiquinone in a reaction that is associated with proton translocation across the membrane.

Number of members: 546

5   [1] Walker JE; Medline: 93110040   The NADH:ubiquinone oxidoreductase (complex I) of respiratory chains." Q Rev Biophys 1992;25:253-324.

978. oxidored_q2. NADH-UBIQUINONE OXIDOREDUCTASE CHAIN 4L (EC 1.6.5.3). ND4L OR NAD4L. Arabidopsis thaliana (Mouse-ear cress). Mitochondrion. OC Eukaryota;

10   Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; Rosidae; eurosids II; Brassicales; Brassicaceae; Arabidopsis.

CATALYTIC ACTIVITY: NADH + UBIQUINONE = NAD(+) + UBIQUINOL.

[1] SEQUENCE FROM N.A. MEDLINE; 93156682. Brandt P., Sunkel S., Unseld M.,

15   Brennicke A., Knoop V.; "The nad4L gene is encoded between exon c of nad5 and orf25 in the Arabidopsis mitochondrial genome."; Mol. Gen. Genet. 236:33-38(1992).

[2] SEQUENCE FROM N.A. STRAIN=CV. COLUMBIA; MEDLINE; 97141919 Unseld M., Marienfeld J.R., Brandt P., Brennicke A.; "The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides."; Nat. Genet. 15:57-61(1997).

20

979. oxidored_q4. Protein name NADH-PLASTOQUINONE OXIDOREDUCTASE CHAIN 3, CHLOROPLAST. Synonym(s)EC 1.6.5.3. Gene name(s)NDHC OR NDH3 From Zea mays (Maize) Encoded on Chloroplast. Taxonomy Eukaryota; Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; Zea.

25   CATALYTIC ACTIVITY: NADH + PLASTOQUINONE = NAD(+) + PLASTOQUINOL.

SIMILARITY: BELONGS TO THE COMPLEX I SUBUNIT 3 FAMILY.

[1] SEQUENCE FROM N.A. MEDLINE; 89281491. Steinmueller K., Ley A.C., Steinmetz

30   A.A., Sayre R.T., Bogorad L.; "Characterization of the ndhC-psbG-ORF157/159 operon of maize plastid DNA and of the cyanobacterium Synechocystis sp. PCC6803."; Mol. Gen. Genet. 216:60-69(1989).

[2] SEQUENCE FROM N.A. MEDLINE; 95395841. Maier R.M., Neckermann K., Igloi G.L., Koessel H.; "Complete sequence of the maize chloroplast genome: gene content,

hotspots of divergence and fine tuning of genetic information by transcript editing."; J. Mol. Biol. 251:614-628(1995).

980. PAC: PAC motif

5   PAC motif occurs C-terminal to a subset of all known PAS motifs. It is proposed to contribute to the PAS domain fold [3]. Number of members: 181

[1] Medline: 97446881 PAS domain S-boxes in archaea, bacteria and sensors for oxygen and redox. Zhulin IB, Taylor BL, Dixon R; Trends Biochem Sci 1997;22:331-333.

10   [2] Medline: 95275818. 1.4 A structure of photoactive yellow protein, a cytosolic photoreceptor: unusual fold, active site, and chromophore. Borgstahl GE, Williams DR, Getzoff ED; Biochemistry 1995;34:6278-6287.

[3] Medline: 98044337. PAS: a multifunctional domain family comes to light. Ponting CP, Aravind L; Curr Biol 1997;7:674-677.

15

981. PARP: Poly(ADP-ribose) polymerase catalytic region.

Poly(ADP-ribose) polymerase catalyses the covalent attachment of ADP-ribose units from NAD+ to itself and to a limited number of other DNA binding proteins, which decreases their affinity for DNA. Poly(ADP-ribose) polymerase is a regulatory component induced by DNA

20   damage.

The carboxyl-terminal region is the most highly conserved region of the protein. Experiments have shown that a carboxyl 40 kDa fragment is still catalytically active [2]. Number of members: 19

25

[1] Medline: 96353841 Structure of the catalytic fragment of poly(AD-ribose) polymerase from chicken. Ruf A, Mennissier de Murcia J, de Murcia G, Schulz GE; Proc Natl Acad Sci U S A 1996;93:7481-7485.

[2] Medline: 93293867 The carboxyl-terminal domain of human poly(ADP-ribose)

30   polymerase. Overproduction in Escherichia coli, large scale purification, and characterization. Simonin F, Hofferer L, Panzeter PL, Muller S, de Murcia G, Althaus FR; J Biol Chem 1993;268:13454-13461.

982. PC_rep: Proteasome/cyclosome repeat

[1] Medline: 97348748 A repetitive sequence in subunits of the 26S proteasome and 20S cyclosome (anaphase-promoting complex). Lupas A, Baumeister W, Hofmann K; Trends Biochem Sci 1997;22:195-196.

Number of members: 112

5

983. Peptidase_M1: Peptidase family M1

Members of this family are aminopeptidases. The members differ widely in specificity, hydrolysing acidic, basic or neutral N-terminal residues. This family includes leukotriene-A4 hydrolase Swiss:P09960, this enzyme also has an aminopeptidase activity [1]. Number of

10   members: 72

[1] Medline: 95405261 Evolutionary families of metallopeptidases. Rawlings ND, Barrett AJ; Meth Enzymol 1995;248:183-228.

15   984. Neutral zinc metallopeptidases, zinc-binding region signature (Peptidase_M8)
PROSITE cross-reference(s) PS00142; ZINC_PROTEASE

The majority of zinc-dependent metallopeptidases (with the notable exception of the carboxypeptidases) share a common pattern of primary structure [1,2,3] in the part of their

20   sequence involved in the binding of zinc, and can be grouped together as a superfamily,known as the metzincins, on the basis of this sequence similarity. They can be classified into a number of distinct families [4,E1] which are listed below along with the proteases which are currently known to belong to these families.

Family M1

25   - Bacterial aminopeptidase N (EC 3.4.11.2) (gene pepN).

- Mammalian aminopeptidase N (EC 3.4.11.2).

- Mammalian glutamyl aminopeptidase (EC 3.4.11.7) (aminopeptidase A). It may play a role in regulating growth and differentiation of early B-lineage cells.

- Yeast aminopeptidase yscII (gene APE2).

30   - Yeast alanine/arginine aminopeptidase (gene AAP1).

- Yeast hypothetical protein YIL137c.

- Leukotriene A-4 hydrolase (EC 3.3.2.6). This enzyme is responsible for the hydrolysis of an epoxide moiety of LTA-4 to form LTB-4; it has been shown that it binds zinc and is capable of peptidase activity.

Family M2

- Angiotensin-converting enzyme (EC 3.4.15.1) (dipeptidyl carboxypeptidase I) (ACE) the enzyme responsible for hydrolyzing angiotensin I to angiotensin II. There are two forms of ACE: a testis-specific isozyme and a somatic isozyme which has two active centers.

5 Family M3

- Thimet oligopeptidase (EC 3.4.24.15), a mammalian enzyme involved in the cytoplasmic degradation of small peptides.

- Neurolysin (EC 3.4.24.16) (also known as mitochondrial oligopeptidase M or microsomal endopeptidase).

10 - Mitochondrial intermediate peptidase precursor (EC 3.4.24.59) (MIP). It is involved the second stage of processing of some proteins imported in the mitochondrion.

- Yeast saccharolysin (EC 3.4.24.37) (proteinase yscD).

- Escherichia coli and related bacteria dipeptidyl carboxypeptidase (EC 3.4.15.5) (gene dcp).

15 - Escherichia coli and related bacteria oligopeptidase A (EC 3.4.24.70) (gene opdA or prlC).

- Yeast hypothetical protein YKL134c.

Family M4

- Thermostable thermolysins (EC 3.4.24.27), and related thermolabile neutral proteases (bacillolysins) (EC 3.4.24.28) from various species of Bacillus.

20 - Pseudolysin (EC 3.4.24.26) from Pseudomonas aeruginosa (gene lasB).

- Extracellular elastase from Staphylococcus epidermidis.

- Extracellular protease prt1 from Erwinia carotovora.

- Extracellular minor protease smp from Serratia marcescens.

- Vibriolysin (EC 3.4.24.25) from various species of Vibrio.

25 - Protease prtA from Listeria monocytogenes.

- Extracellular proteinase proA from Legionella pneumophila.

Family M5

- Mycolysin (EC 3.4.24.31) from Streptomyces cacaoi.

30

Family M6

- Immune inhibitor A from Bacillus thuringiensis (gene ina). Ina degrades two classes of insect antibacterial proteins, attacins and cecropins.

Family M7

- Streptomyces extracellular small neutral proteases

Family M8

5   - Leishmanolysin (EC 3.4.24.36) (surface glycoprotein gp63), a cell surface protease from various species of Leishmania.

Family M9

- Microbial collagenase (EC 3.4.24.3) from Clostridium perfringens and Vibrio
10   alginolyticus.

Family M10A

- Serralysin (EC 3.4.24.40), an extracellular metalloprotease from Serratia.
- Alkaline metalloproteinase from Pseudomonas aeruginosa (gene aprA).
15   - Secreted proteases A, B, C and G from Erwinia chrysanthemi.
- Yeast hypothetical protein YIL108w.

Family M10B

- Mammalian extracellular matrix metalloproteinases (known as matrixins) [5]: MMP-1 (EC
20   3.4.24.7) (interstitial collagenase), MMP-2 (EC 3.4.24.24) (72 Kd gelatinase), MMP-9 (EC 3.4.24.35) (92 Kd gelatinase), MMP-7 (EC 3.4.24.23) (matrylisin), MMP-8 (EC 3.4.24.34) (neutrophil collagenase), MMP-3 (EC 3.4.24.17) (stromelysin-1), MMP-10 (EC 3.4.24.22) (stromelysin-2), and MMP-11 (stromelysin-3), MMP-12 (EC 3.4.24.65) (macrophage metalloelastase).
25   - Sea urchin hatching enzyme (envelysin) (EC 3.4.24.12). A proteas that allows the embryo to digest the protective envelope derived from the egg extracellular matrix.
- Soybean metalloendoproteinase 1.

Family M11

30   - Chlamydomonas reinhardtii gamete lytic enzyme (GLE).

Family M12A

- Astacin (EC 3.4.24.21), a crayfish endoprotease.
- Meprin A (EC 3.4.24.18), a mammalian kidney and intestinal brush border

metalloendopeptidase.

- Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation and which expresses metalloendopeptidase activity. The Drosophila homolog of BMP-1 is the dorsal-ventral patterning protein tolloid.

5   - Blastula protease 10 (BP10) from Paracentrotus lividus and the related protein SpAN from Strongylocentrotus purpuratus.

- Caenorhabditis elegans protein toh-2.

- Caenorhabditis elegans hypothetical protein F42A10.8.

- Choriolysins L and H (EC 3.4.24.67) (also known as embryonic hatching proteins LCE

10   and HCE) from the fish Oryzias lapides. These proteases participates in the breakdown of the egg envelope, which is derived from the egg extracellular matrix, at the time of hatching.

Family M12B

15   - Snake venom metalloproteinases [6]. This subfamily mostly groups proteases that act in hemorrhage. Examples are: adamalysin II (EC 3.4.24.46), atrolysin C/D   (EC 3.4.24.42), atrolysin E (EC 3.4.24.44), fibrolase (EC 3.4.24.72), trimerelysin I (EC 3.4.25.52) and II (EC 3.4.25.53).

- Mouse cell surface antigen MS2.

20

Family M13

- Mammalian neprilysin (EC 3.4.24.11) (neutral endopeptidase) (NEP).

- Endothelin-converting enzyme 1 (EC 3.4.24.71) (ECE-1), which process the precursor of endothelin to release the active peptide.

25   - Kell blood group glycoprotein, a major antigenic protein of erythrocytes. The Kell protein is very probably a zinc endopeptidase.

- Peptidase O from Lactococcus lactis (gene pepO).

Family M27

30   - Clostridial neurotoxins, including tetanus toxin (TeTx) and the various botulinum toxins (BoNT). These toxins are zinc proteases that block neurotransmitter release by proteolytic cleavage of synaptic proteins such as synaptobrevins, syntaxin and SNAP-25 [7,8].

Family M30

- Staphylococcus hyicus neutral metalloprotease.

Family M32

5 - Thermostable carboxypeptidase 1 (EC 3.4.17.19) (carboxypeptidase Taq), an enzyme from Thermus aquaticus which is most active at high temperature.

Family M34

- Lethal factor (LF) from Bacillus anthracis, one of the three proteins composing the

10 anthrax toxin.

Family M35

- Deuterolysin (EC 3.4.24.39) from Penicillium citrinum and related proteases from various species of Aspergillus.

15

Family M36

- Extracellular elastinolytic metalloproteinases from Aspergillus.

From the tertiary structure of thermolysin, the position of the residues acting as zinc

20 ligands and those involved in the catalytic activity are known. Two of the zinc ligands are histidines which are very close together in the sequence; C-terminal to the first histidine is a glutamic acid residue which acts as a nucleophile and promotes the attack of a water molecule on the carbonyl carbon of the substrate. A signature pattern which includes the two histidine and the glutamic acid residues is sufficient to detect this superfamily of

25 proteins.

Consensus pattern[GSTALIVN]-x(2)-H-E-[LIVMFYW]-{DEHRKP}-H-x-[LIVMFYWGSPQ]

[The two H's are zinc ligands] [E is the active site residue]

30   Sequences known to belong to this class detected by the patternALL, except

for members of families M5, M7 amd M11.

Other sequence(s) detected in SWISS-PROT57; including Neurospora crassa

conidiation-specific protein 13 which could be a zinc-protease.

[1]Jongeneel C.V., Bouvier J., Bairoch A. FEBS Lett. 242:211-214(1989).

[2]Murphy G.J.P., Murphy G., Reynolds J.J. FEBS Lett. 289:4-7(1991).

[3]Bode W., Grams F., Reinemer P., Gomis-Rueth F.-X., Baumann U., McKay D.B., Stoecker W. Zoology 99:237-246(1996).

[4]Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).

5   [5]Woessner J. Jr. FASEB J. 5:2145-2154(1991).

[6]Hite L.A., Fox J.W., Bjarnason J.B. Biol. Chem. Hoppe-Seyler 373:381-385(1992).

[7]Montecucco C., Schiavo G. Trends Biochem. Sci. 18:324-327(1993).

[8]Niemann H., Blasi J., Jahn R. Trends Cell Biol. 4:179-185(1994).

10   985. PHO4: Phosphate transporter family

This family includes PHO-4 from Neurospora crassa which is a is a Na(+)-phosphate symporter [1]. This family also contains the leukemia virus receptor Swiss:Q08344. Number of members: 41

15   [1] Medline: 95249577 Repressible cation-phosphate symporters in Neurospora crassa. Versaw WK, Metzenberg RL; Proc Natl Acad Sci U S A 1995;92:3884-3887.

986. Photosynthetic reaction center proteins signature (photoRC)
PROSITE cross-reference(s): PS00244; REACTION_CENTER

20

In the photosynthetic reaction center of purple bacteria, two homologous integral membrane proteins, L(ight) and M(edium), are known to be essential to the light-mediated water-splitting process.   In the photosystem II of eukaryotic chloroplasts two related proteins are involved: the D1 (psbA) and D2 proteins (psbD).   These four types of protein

25   probably evolved from a common ancestor [see 1,2 for recent reviews].

A signature pattern was developed which include two conserved histidine residues. In L and M chains, the first histidine is a ligand of the magnesium ion of the special pair bacteriochlorophyll, the second is a ligand of a ferrous non-heme iron atom. In photosystem

30   II these two histidines are thought to play a similar role.

Consensus pattern[NQH]-x(4)-P-x-H-x(2)-[SAG]-x(11)-[SAGC]-x-H-[SAG](2)
[The first H is a magnesium ligand] [The second H is a iron ligand]
Sequences known to belong to this class detected by the patternALL, except

for broad bean psbA which has Gln instead of the second His.

[1]Michel H., Deisenhofer J. Biochemistry 27:1-7(1988).

[2]Barber J. Trends Biochem. Sci. 12:321-326(1987).

987. phytochrome: Phytochrome region

This family contains a region specific to phytochrome proteins. Number of members:
145

988. PI3K_C2: C2 domain

Phosphoinositide 3-kinase region postulated to contain a C2 domain. Outlier of C2 family.
Number of members: 39

[1] Medline: 97388296 Using structure to define the function of phosphoinositide 3-kinase family members. Domin J, Waterfield MD; FEBS Lett 1997;410:91-95.

[2] Medline: 97398940 Phosphoinositide 3-kinases: a conserved family of signal transducers. Vanhaesebroeck B, Leevers SJ, Panayotou G, Waterfield MD; Trends Biochem Sci 1997;22:267-272.

989. PI3Ka: Phosphoinositide 3-kinase family, accessory domain (PIK domain)

PIK domain is conserved in all PI3 and PI4-kinases. Its role is unclear but it has been suggested [2] to be involved in substrate presentation.

Number of members: 47

[1] Medline: 97388296 Using structure to define the function of phosphoinositide 3-kinase family members. Domin J, Waterfield MD; FEBS Lett 1997;410:91-95.

[2] Medline: 94069320 Phosphatidylinositol 4-kinase: gene structure and requirement for yeast cell viability. Flanagan CA, Schnieders EA, Emerick AW, Kunisawa R, Admon A, Thorner J; Science 1993;262:1444-1448.

990. P-II protein signatures

PROSITE cross-reference(s): PS00496; PII_GLNB_UMP, PS00638; PII_GLNB_CTER

The P-II protein (gene glnB) is a bacterial protein important for the control of glutamine synthetase [1,2,3]. In nitrogen-limiting conditions, when the ratio of glutamine to 2-ketoglutarate decreases, P-II is uridylylated on a tyrosine residue to form P-II-UMP. P-II-UMP allows the deadenylation of glutamine synthetase (GS), thus activating the enzyme.

5 Conversely, in nitrogen excess, P-II-UMP is deuridylated and then promotes the adenylation of GS. P-II also indirectly controls the transcription of the GS gene (glnA) by preventing NR-II (ntrB) to phosphorylate NR-I (ntrC) which is the transcriptional activator of glnA. Once P-II is uridylylated, these events are reversed.

10 P-II is a protein of about 110 amino acid residues extremely well conserved. The tyrosine which is urydylated is located in the central part of the protein.

In cyanobacteria, P-II seems to be phosphorylated on a serine residue rather than being urydylated.

15

In methanogenic archaebacteria, the nitrogenase iron protein gene (nifH) is followed by two open reading frames highly similar to the eubacterial P-II protein [4]. These proteins could be involved in the regulation of nitrogen fixation.

20 In the red alga, Porphyra purpurea, there is a glnB homolog encoded in the chloroplast genome.

Other proteins highly similar to glnB are:

25 - Bacillus subtilis protein nrgB [5].
- Escherichia coli hypothetical protein ybaI [6].

Two signature patterns were developed for P-II protein. The first one is a conserved stretch (in eubacteria) of six residues which contains the urydylated tyrosine, the other
30 is derived from a conserved region in the C-terminal part of the P-II protein.

Consensus patternY-[KR]-G-[AS]-[AE]-Y [The second Y is uridylated]
Sequences known to belong to this class detected by the patternALL glnB's
from eubacteria.

Consensus pattern[ST]-x(3)-G-[DY]-G-[KR]-[IV]-[FW]-[LIVM]-x(2)-[LIVM]

[1]Magasanik B. Biochimie 71:1005-1012(1989).

[2]Holtel A., Merrick M. Mol. Gen. Genet. 215:134-138(1988).

[3]Cheah E., Carr P.D., Suffolk P.M., Vasuvedan S.G., Dixon N.E., Ollis D.L. Structure 2:981-990(1994).

[4]Sibold L., Henriquet M., Possot O., Aubert J.-P. Res. Microbiol. 142:5-12(1991).

[5]Wray L.V. Jr., Atkinson M.R., Fisher S.H. J. Bacteriol. 176:108-114(1994).

[6]Allikmets R., Gerrard B.C., Court D., Dean M.C. Gene 136:231-236(1993).

991. PIP5K: Phosphatidylinositol-4-phosphate 5-Kinase

This family contains a region from the common kinase core found in the type I phosphatidylinositol-4-phosphate 5-kinase (PIP5K) family as described in [1]. The family consists of various type I, II and III PIP5K enzymes. PIP5K catalyses the formation of phosphoinositol-4,5-bisphosphate via the phosphorylation of phosphatidylinositol-4-phosphate a precursor in the phosphinositide signaling pathway. Number of members: 33

[1] Medline: 98204859. Type I phosphatidylinositol-4-phosphate 5-kinases. Cloning of the third isoform and deletion/substitution analysis of members of this novel lipid kinase family. Ishihara H, Shibasaki Y, Kizuki N, Wada T, Yazaki Y, Asano T, Oka Y; J Biol Chem 1998;273:8741-8748.

[2] Medline: 97115834 Type I phosphatidylinositol-4-phosphate 5-kinases are distinct members of this novel lipid kinase family. Loijens JC, Anderson RA; J Biol Chem 1996 20;271:32937-32943.

992. PolyA_pol: Poly A polymerase family

This family includes nucleic acid independent RNA polymerases, such as Poly(A) polymerase, which adds the poly (A) tail to mRNA EC:2.7.7.19. This family also includes the tRNA nucleotidyltransferase that adds the CCA to the 3' of the tRNA EC:2.7.7.25. Number of members: 31

[1] Medline: 93066242 Identification of the gene for an Escherichia coli poly(A) polymerase. Cao GJ, Sarkar N; Proc Natl Acad Sci U S A 1992;89:10380-10384.

993. Photosystem I psaA and psaB proteins signature (psaA_psaB)

PROSITE cross-reference(s)PS00419; PHOTOSYSTEM_I_PSAAB

Photosystem I (PSI) [1] is an integral membrane protein complex that uses light energy to
5   mediate electron transfer from plastocyanin to ferredoxin. PSI is found in the chloroplast
of plants and cyanobacteria. The electron transfer components of the reaction center of
PSI are a primary electron donor P-700 (chlorophyll dimer) and five electron acceptors: A0
(chlorophyll), A1 (a phylloquinone) and three 4Fe-4S iron-sulfur centers: Fx, Fa, and Fb.

10   PsaA and psaB, two closely related proteins, are involved in the binding of P700, A0, A1,
and Fx. psaA and psaB are both integral membrane proteins of 730 to 750 amino acids that
seem to contain 11 transmembrane segments. The Fx 4Fe-4S iron-sulfur center is bound by
four cysteines; two of these cysteines are provided by the psaA protein and the two others
by psaB. The two cysteines in both proteins are proximal and located in a loop between
15   the ninth and tenth transmembrane segments. A leucine zipper motif seems to be present [2]
downstream of the cysteines and could contribute to dimerization of psaA/psaB.

The signature pattern for these proteins is based on the perfectly conserved region that
includes the two iron-sulfur binding cysteines.

20   Consensus patternC-D-G-P-G-R-G-G-T-C [The two C's bind the iron-sulfur center]

[1]Golbeck J.H. Biochim. Biophys. Acta 895:167-204(1987).
[2]Webber A.N., Malkin R. FEBS Lett. 264:1-14(1990).

25   994. PSBH: Photosystem II 10 kDa phosphoprotein
This protein is phosphorylated in a light dependent reaction.
Number of members: 20

995. PsbJ

30   This family consists of the photosystem II reaction center protein PsbJ from plants and
Cyanobacteria. In Synechocystis sp. PCC 6803 PsbJ regulates the number of photosystem II
centers in thylakoid membranes, it is a predicted 4kDa protein with one membrane spanning
domain [1]. Number of members: 20

[1] Medline: 93131892. Genetic and immunological analyses of the cyanobacterium Synechocystis sp. PCC 6803 show that the protein encoded by the psbJ gene regulates the number of photosystem II centers in thylakoid membranes. Lind LK, Shukla VK, Nyhus KJ, Pakrasi HB; J Biol Chem 1993;268:1575-1579.

5

996. PSBT: Photosystem II reaction centre T protein

The exact function of this protein is unknown. It probably consists of a single transmembrane spanning helix. The Swiss:P37256 protein, appears to be (i) a novel photosystem II subunit and (ii) required for maintaining optimal photosystem II activity under adverse growth

10 conditions [1]. Number of members: 17

[1] Medline: 94298765. The chloroplast ycf8 open reading frame encodes a photosystem II polypeptide which maintains photosynthetic activity under adverse growth conditions. Monod C, Takahashi Y, Goldschmidt-Clermont M, Rochaix JD; EMBO J

15 1994;13:2747-2754.

997. PSI_8. PHOTOSYSTEM I REACTION CENTRE SUBUNIT VIII. Synonym(s)PSI-I. Gene name(s)PSAI. From Hordeum vulgare (Barley). Encoded on Chloroplast. Taxonomy Eukaryota; Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta;

20 Liliopsida; Poales; Poaceae; Hordeum.

MAY HELP IN THE ORGANIZATION OF THE PSAL SUBUNIT. BELONGS TO THE PSAI FAMILY.

[1] SEQUENCE FROM N.A. MEDLINE; 90036933. Scheller H.V., Okkels J.S., Hoej P.B.,

25 Svendsen I., Roepstorff P., Moeller B.L.; "The primary structure of a 4.0-kDa photosystem I polypeptide encoded by the chloroplast psaI gene."; J. Biol. Chem. 264:18402-18406(1989).

998. PSI_PsaJ: Photosystem I reaction centre subunit IX / PsaJ

This family consists of the photosystem I reaction centre subunit IX or PsaJ from various

30 organisms including Synechocystis sp. (strain pcc 6803), Pinus thunbergii (green pine) and Zea mays (maize). PsaJ Swiss:P19443 is a small 4.4kDa, chloroplastal encoded, hydrophobic subunit of the photosystem I reaction complex its function is not yet fully understood [1]. PsaJ can be cross-linked to PsaF Swiss:P12356 and has a single predicted transmembrane

domain it has a proposed role in maintaing PsaF in the correct orientation to allow for fast electron transfer from soluble donor proteins to P700+ [1]. Number of members: 18

[1] Medline: 99238330. A large fraction of PsaF is nonfunctional in photosystem I complexes
5 lacking the PsaJ subunit. Fischer N, Boudreau E, Hippler M, Drepper F, Haehnel W, Rochaix JD; Biochemistry 1999;38:5546-5552.

[2] Medline: 93252282. Genes encoding eleven subunits of photosystem I from the thermophilic cyanobacterium Synechococcus sp. Muhlenhoff U, Haehnel W, Witt H, Herrmann RG; Gene 1993;127:71-78.

10

999. PSII. Protein namePHOTOSYSTEM II P680 CHLOROPHYLL A APOPROTEIN. Synonym(s)CP-47 PROTEIN. Gene name(s)PSBB. From Hordeum vulgare (Barley), Encoded on Chloroplast. Taxonomy Eukaryota; Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; Hordeum.

15   FUNCTION: THIS PROTEIN CONJUGATES WITH CHLOROPHYLL & CATALYZES THE PRIMARY LIGHT-INDUCED PHOTOCHEMICAL PROCESSES OF PHOTOSYSTEM II. SUBCELLULAR LOCATION: CHLOROPLAST THYLAKOID MEMBRANE. SIMILARITY: BELONGS TO THE PSBB / PSBC FAMILY.

20 [1] SEQUENCE FROM N.A. STRAIN=CV. SABARLIS; MEDLINE; 89240047. Andreeva A.V., Buryakova A.A., Reverdatto S.V., Chakhmakhcheva O.G., Efimov V.A.; "Nucleotide sequence of the 5.2 kbp barley chloroplast DNA fragment, containing psbB-psbH-petB-petD gene cluster."; Nucleic Acids Res. 17:2859-2860(1989).

[2] SEQUENCE FROM N.A. STRAIN=CV. SABARLIS; MEDLINE; 92207253. Efimov
25 V.A., Andreeva A.V., Reverdatto S.V., Chakhmakhcheva O.G.; "Photosystem II of rye. Nucleotide sequence of the psbB, psbC, psbE, psbF, psbH genes of rye and chloroplast DNA regions adjacent to them."; Bioorg. Khim. 17:1369-1385(1991).

[3] SEQUENCE OF 411-420. Hinz U.G.; "Isolation of the photosystem II reaction center complex from barley. Characterization by cicular dichroism spectroscopy and amino acid
30 sequencing."; Carlsberg Res. Commun. 50:285-298(1985).

1000. QRPTase. Quinolinate phosphoribosyl transferase.
Quinolinate phosphoribosyl transferase (QPRTase) or nicotinate-nucleotide pyrophosphorylase EC:2.4.2.19 is involved in the de novo synthesis of NAD in both

prokaryotes and eukaryotes. It catalyses the reaction of quinolinic acid with 5-phosphoribosyl-1-pyrophosphate (PRPP) in the presence of Mg2+ to give rise to nicotinic acid mononucleotide (NaMN), pyrophosphate and carbon dioxide [1,2]. Number of members: 26.

[1]Medline: 97169443. A new function for a common fold: the crystal structure of quinolinic acid phosphoribosyltransferase. Eads JC, Ozturk D, Wexler TB, Grubmeyer C, Sacchettini JC; Structure 1997;5:47-58.

[2]Medline: 96139309. The sequencing expression, purification, and steady-state kinetic analysis of quinolinate phosphoribosyl transferase from Escherichia coli. Bhatia R, Calvo KC; Arch Biochem Biophys 1996;325:270-278.

1001. R3H domain

The name of the R3H domain comes from the characteristic spacing of the most conserved arginine and histidine residues. The function of the domain is predicted to be binding ssDNA. Number of members: 28

[1]Medline: 99003905 The R3H motif: a domain that binds single-stranded nucleic acids. Grishin NV; Trends Biochem Sci 1998;23:329-330.

1002. recF protein signatures (RecF)

The prokaryotic protein recF [1,2] is a single-stranded DNA-binding protein which also probably binds ATP. RecF is involved in DNA metabolism; it is required for recombinational DNA repair and for induction of the SOS response. RecF is a protein of about 350 to 370 amino acid residues; there is a conserved ATP-binding site motif 'A' (P-loop) in the N-terminal section of the protein as well as two other conserved regions, one located in the central section, and the other in the C-terminal section. Signature patterns were derived from these two regions.

Consensus pattern [LIVM]-x(4)-[LIF]-x(6)-[LIF]-[LVF]-x-[GE]-[GSTAD]-[PA]- x(2)-R-R-x-[FYW]-[LIVMF]-D Sequences known to belong to this class detected by the pattern ALL.

779

Consensus pattern[LIVMFY](2)-x-D-x(2,3)-[SA]-[EH]-L-D-x(2)-[KRH]-x(3)-L Sequences known to belong to this class detected by the patternALL, except for T. palidum recF.

[1] Sandler S.J., Chackerian B., Li J.T., Clark A.J. Nucleic Acids Res. 20:839-845(1992).

5   [2] Alonso J.C., Fisher L.M.; Mol. Gen. Genet. 246:680-686(1995).

1003. RibD C-terminal domain (RibD_C)

The function of this domain is not known, but it is thought to be involved in riboflavin

10   biosynthesis. This domain is found in the C terminus of RibD/RibG Swiss:P25539, in combination with dCMP_cyt_deam, as well as in isolation in some archaebacterial proteins Swiss:P95872.

Number of members: 21

15   1004. Ribosomal protein L16 signatures (Ribosomal_L16)

Ribosomal protein L16 is one of the proteins from the large ribosomal subunit. In Escherichia coli, L16 is known to bind directly the 23S rRNA and to be located at the A site of the peptidyltransferase center. It belongs to a family of ribosomal proteins which, on the basis of

20   sequence similarities [1], groups:
    - Eubacterial L16.
    - Algal and plant chloroplast L16.
    - Cyanelle L16.
    - Plant mitochondrial L16.

25   L16 is a protein of 133 to 185 amino-acid residues. As signature patterns, we selected two conserved regions in the central section of these proteins.

Consensus pattern [KR](2)-x-[GSAC]-[KRQVA]-[LIVM]-W-[LIVM]-[KR]-[LIVM]-[LFY]-[AP] Sequences known to belong to this class detected by the pattern ALL.

30

Consensus patternR-M-G-x-[GR]-K-G-x(4)-[FWKR] Sequences known to belong to this class detected by the patternALL.

[1] Otaka E., Hashimoto T., Mizuta K., Suzuki K. Protein Seq. Data Anal. 5:301-313(1993).

1005. Ribosomal protein L32e signature (Ribosomal_L32E)

A number of eukaryotic and archaebacterial ribosomal proteins can be grouped on the basis
of sequence similarities. One of these families consists of:

- Mammalian L32 [1].
- Drosophila RP49 [2].
- Trichoderma harzianum L32 [3].
- Yeast L32e (YBL092w).
- Archaebacterial L32e [4].

These proteins have 135 to 240 amino-acid residues. As a signature pattern, a stretch of about
20 residues located in the N-terminal part of these proteins was seleced.

Consensus patternF-x-R-x(4)-[KR]-x(2)-[KR]-[LIVMF]-x(3,5)-W-R-[KR]-x(2)-G Sequences
known to belong to this class detected by the pattern ALL.

[1] Jacks C.M., Powaser C.B., Hackett P.B. Gene 74:565-570(1988).

[2] Aguade M. Mol. Biol. Evol. 5:433-441(1988).

[3] Lora J.M., Garcia I., Benitez T., Llobell A., Pintor-Toro J.A. Nucleic Acids Res.
21:3319-3319(1993).

[4] Arndt E., Scholzen T., Kroemer W., Hatakeyama T., Kimura M. Biochimie 73:657-
668(1991).

1006. (Ribosomal_S3) Ribosomal protein S3 signature

PROSITE: PDOC00474. PROSITE cross-reference(s) PS00548; RIBOSOMAL_S3

Ribosomal protein S3 is one of the proteins from the small ribosomal subunit.
In Escherichia coli, S3 is known to be involved in the binding of initiator Met-tRNA. It
belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1],
groups:

-Eubacterial S3.

-Algal and plant chloroplast S3.

-Cyanelle S3.

-Archaebacterial S3.

-Plant mitochondrial S3.

-Vertebrate S3.

-Insect S3.

-Caenorhabditis elegans S3 (C23G10.3).

-Yeast S3 (Rp13).

5   S3 is a protein of 209 to 559 amino-acid residues. A conserved region located in the C-terminal section was selected as a signature pattern.

Consensus pattern[GSTA]-[KR]-x(6)-G-x-[LIVMT]-x(2)-[NQSCH]-x(1,3)-[LIVFCA]-x(3)-[LIV]-[DENQ]-x(7)-[LMT]-x(2)-G-x(2)-[GS]. Sequences known to belong to this class

10 detected by the patternALL, except for some mitochondrial S3.

[1]Otaka E., Hashimoto T., Mizuta K. Protein Seq. Data Anal. 5:285-300(1993).

1007. RimM - RimM

15 The RimM protein is essential for efficient processing of 16S rRNA [1]. The RimM protein was shown to have affinity for free ribosomal 30S subunits but not for 30S subunits in the 70S ribosomes [1]. Number of members: 14.

[1]Medline: 98083058. RimM and RbfA are essential for efficient processing of 16S rRNA in

20 Escherichia coli. Bylund GO, Wipemo LC, Lundberg LA, Wikstrom PM; J Bacteriol 1998;180:73-82.

1008. RNA_pol_A - RNA polymerase alpha subunit

-!- RNA polymerases catalyse the DNA dependent polymerisation of RNA. Prokaryotes

25 contain a single RNA polymerase compared to three in eukaryotes (not including mitochondrial and chloroplast polymerases).

-!- Members of this family include: A subunit from eukaryotes, gamma subunit from cyanobacteria, beta' subunit from eubacteria, A' subunit from archaebacteria, B" from chloroplasts. Number of members: 139.

30

[1]Medline: 97066998. Structural modules of the large subunits of RNA polymerase. Introducing archaebacterial and chloroplast split sites in the beta and beta' subunits of Escherichia coli RNA polymerase. Severinov K, Mustaev A, Kukarin A, Muzzin O, Bass I, Darst SA, Goldfarb A; J Biol Chem 1996;271:27969-27974.

1009. RuBisCO_large - Ribulose bisphosphate carboxylase large chain active site

PROSITE: PDOC00142; PROSITE cross-reference(s) PS00157; RUBISCO_LARGE

Ribulose bisphosphate carboxylase (EC 4.1.1.39) (RuBisCO) [1,2] catalyzes the

5   initial step in Calvin's reductive pentose phosphate cycle in plants as well as purple and green

bacteria. It consists of a large catalytic unit and a small subunit of undetermined function. In

plants, the large subunit is coded by the chloroplastic genome while the small subunit is

encoded in the nuclear genome. Molecular activation of RuBisCO by CO_2 involves the

formation of a carbamate with the epsilon-amino group of a conserved lysine residue. This

10   carbamate is stabilized by a magnesium ion. One of the ligands of the magnesium ion is an

aspartic acid residue close to the active site lysine [3]. A pattern was developed which

includes both the active site residue and the metal ligand, and which is specific to RuBisCO

large chains.

15   Consensus patternG-x-[DN]-F-x-K-x-D-E [K is the active site residue] [The second D is a

magnesium ligand]. Sequences known to belong to this class detected by the patternALL,

except for Cheilopleuria biscuspis RuBisCO.

[1]Miziorko H.M., Lorimer G.H. Annu. Rev. Biochem. 52:507-535(1983).

20   [2]Akazawa T., Takabe T., Kobayashi H. Trends Biochem. Sci. 9:380-383(1984).

[3]Andersson I., Knight S., Schneider G., Lindqvist Y., Lundqvist T., Branden C.-I., Lorimer

G.H. Nature 337:229-234(1989).

1010. Rve - Integrase core domain

25   Integrase mediates integration of a DNA copy of the viral genome into the host chromosome.

Integrase is composed of three domains. The amino-terminal domain is a zinc binding

domain Integrase_Zn. This domain is the central catalytic domain. The carboxyl terminal

domain that is a non-specific DNA binding domain integrase. The catalytic domain acts as an

endonuclease when two nucleotides are removed from the 3' ends of the blunt-ended viral

30   DNA made by reverse transcription. This domain also catalyses the DNA strand transfer

reaction of the 3' ends of the viral DNA to the 5' ends of the integration site [1]. Number of

members: 694.

[1]Medline: 95099322. Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. Dyda F, Hickman AB, Jenkins TM, Engelman A, Craigie R, Davies DR; Science 1994;266:1981-1986.

5   1011. (SBP_bac_3) Bacterial extracellular solute-binding proteins, family 3 signature
PROSITE: PDOC00798. PROSITE cross-reference(s) PS01039; SBP_BACTERIAL_3

Bacterial high affinity transport systems are involved in active transport of solutes across the cytoplasmic membrane. The protein components of these traffic systems include one or two transmembrane protein components, one or two membrane-associated ATP-

10   binding proteins (ABC transporters; see <PDOC00185>) and a high affinity periplasmic solute-binding protein. The later are thought to bind the substrate in the vicinity of the inner membrane, and to transfer it to a complex of inner membrane proteins for concentration into the cytoplasm.

In gram-positive bacteria which are surrounded by a single membrane and have

15   therefore no periplasmic region the equivalent proteins are bound to the membrane via an N-terminal lipid anchor. These homolog proteins do not play an integral role in the transport process per se, but probably serve as receptors to trigger or initiate translocation of the solute throught the membrane by binding to external sites of the integral membrane proteins of the efflux system.

20   In addition at least some solute-binding proteins function in the initiation of sensory transduction pathways.

On the basis of sequence similarities, the vast majority of these solute-binding proteins can be grouped [1] into eight families of clusters, which generally correlate with the nature of the solute bound.

25   Family 3 groups together specific amino acids and opine-binding periplasmic proteins and a periplasmic homolog with catalytic activity:
-Histidine-binding protein (gene hisJ) of Escherichia coli and related bacteria. An homologous lipoprotein exists in Neisseria gonorrhoeae.
-Lysine/arginine/ornithine-binding proteins (LAO) (gene argT) of Escherichia coli and

30   related bacteria are involved in the same transport system than hisJ. Both solute-binding proteins interact with a common membrane-bound receptor hisP of the binding protein dependent transport system HisQMP.
-Glutamine-binding proteins (gene glnH) of Escherichia coli and Bacillus stearothermophilus.

-Glutamate-binding protein (gene gluB) of Corynebacterium glutamicum.

-Arginine-binding proteins artI and artJ of Escherichia coli.

-Nopaline-binding protein (gene nocT) from Agrobacterium tumefaciens.

-Octopine-binding protein (gene occT) from Agrobacterium tumefaciens.

5   -Major cell-binding factor (CBF1) (gene: peb1A) from Campylobacter jejuni.

-Bacteroides nodosus protein aabA.

-Cyclohexadienyl/arogenate dehydratase of Pseudomonas aeruginosa, a periplasmic enzyme which forms an alternative pathway for phenylalanine biosynthesis.

-Escherichia coli protein fliY.

10   -Vibrio harveyi protein patH.

-Escherichia coli hypothetical protein ydhW.

-Bacillus subtilis hypothetical protein yckB.

-Bacillus subtilis hypothetical protein yckK.

15   The signature pattern is located near the N-terminus of the mature proteins.

Consensus patternG-[FYIL]-[DE]-[LIVMT]-[DE]-[LIVMF]-x(3)-[LIVMA]-[VAGC]-x(2)-[LIVMAGN]

Sequences known to belong to this class detected by the patternALL.

20   [1]Tam R., Saier M.H. Jr. Microbiol. Rev. 57:320-346(1993).

1012. Sec7 - Sec7 domain

The Sec7 domain is a guanine-nucleotide-exchange-factor (GEF)for the arf family [2]. Number of members: 32.

25

[1]Medline: 98169075. Structure of the Sec7 domain of the Arf exchange factor. ARNO. Cherfils J, Menetrey J, Mathieu M, Le Bras G, Robineau S, Beraud-Dufour S, Antonny B, Chardin P; Nature 1998;392:101-105.

[2]Medline: 97100951. A human exchange factor for ARF contains Sec7- and pleckstrin-
30   homology domains. Chardin P, Paris S, Antonny B, Robineau S, Beraud-Dufour S, Jackson CL, Chabre M. Nature 1996;384:481-484.

1013. SecA_protein. SecA protein, amino terminal region

SecA protein binds to the plasma membrane where it interacts with proOmpA to support translocation of proOmpA through the membrane. SecA protein achieves this translocation, in association with SecY protein, in an ATP dependent manner. SecA possesses the ATPase activity. The carboxyl terminus has similarity with the helicase carboxyl terminus. See

5   Ribosomal_L5. Number of members: 45.

[1]Medline: 98309858. Amino-terminal region of SecA is involved in the function of SecG for protein translocation into Escherichia coli membrane vesicles. Mori H, Sugiyama H, Yamanaka M, Sato K, Tagaya M, Mizushima S; J Biochem (Tokyo) 1998;124:122-129.

10   [2]Medline: 89251629. SecA protein hydrolyzes ATP and is an essential component of the protein translocation ATPase of Escherichia coli. Lill R, Cunningham K, Brundage LA, Ito K, Oliver D, Wickner W; EMBO J 1989;8:961-966.

1014. Seedstore_2S - 2S seed storage family

15   Members of this family are composed of two chains (both included in the alignment), these are co-translated and later cleaved. The two chains are disulphide linked together. Number of members: 27.

[1]Medline: 97121264. 1H NMR assignment and global fold of napin BnIb, a representative

20   2S albumin seed protein. Rico M, Bruix M, Gonzalez C, Monsalve RI, Rodriguez R; Biochemistry 1996;35:15672-15682.

1015. Smr - Smr domain

This family includes the Smr (Small MutS Related) proteins, and the C-terminal region of the

25   MutS2 protein. It has been suggested that this domain interacts with the MutS1 Swiss:P23909 protein in the case of Smr proteins and with the N-terminal MutS related region of MutS2 Swiss:P94545 [1]. Number of members: 14.

[1]Medline: 10431172. Smr: a bacterial and eukaryotic homologue of the C-terminal region

30   of the MutS2 family. Moreira D, Philippe H; Trends Biochem Sci 1999;24:298-300.

1016. (SSF) Sodium:solute symporter family signatures and profile
PROSITE: PDOC00429. PROSITE cross-reference(s)PS00456; NA_SOLUT_SYMP_1
PS00457; NA_SOLUT_SYMP_2 PS50283; NA_SOLUTE_SYMP_3

It has been shown [1,2] that integral membrane proteins that mediate the intake of a wide variety of molecules with the concomitant uptake of sodium ions (sodium symporters) can be grouped, on the basis of sequence and functional similarities into a number of distinct families. One of these families is known as the sodium:solute symporter family (SSF) and

5 currently consists of the following proteins:

-Mammalian Na+/glucose co-transporter.

-Mammalian Na+/myo-inositol co-transporter.

-Mammalian Na+/nucleoside co-transporter.

-Mammalian Na+/neutral amino acid co-transporter.

10 -Escherichia coli Na+/proline symporter (gene putP).

-Escherichia coli Na+/pantothenate symporter (gene panF).

-Escherichia coli hypothetical protein yidK.

-Escherichia coli hypothetical protein yjcG.

-Bacillus subtilis hypothetical protein ywcA (ipa-31R).

15 These integral membrane proteins are predicted to comprise at least ten membrane spanning domains. Two conserved regions were selected as signature patterns; the first one is located in the fourth transmembrane region and the second one in a loop between two transmembrane regions in the C-terminal part of these proteins.

20 Consensus pattern[GS]-x(2)-[LIY]-x(3)-[LIVMFYWSTAG](10)-[LIY]-[TAV]-x(2)-G-G-[LMF]-x-[SAP]. Sequences known to belong to this class detected by the patternALL. Consensus pattern[GAST]-[LIVM]-x(3)-[KR]-x(4)-G-A-x(2)-[GAS]-[LIVMGS]-[LIVMW]-[LIVMGAT]-G-x-[LIVMGA] Sequences known to belong to this class detected by the patternALL, except for E.coli yidK.

25 Note this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so.

[1]Reizer J., Reizer A., Saier M.H. Jr. Res. Microbiol. 141:1069-1072(1991).

30 [2]Reizer J., Reizer A., Saier M.H. Jr. Biochim. Biophys. Acta 1197:133-136(1994).

1017. SurE - Survival protein SurE

E. coli cells with the surE gene disrupted are found to survive poorly in stationary phase [1]. It is suggested that SurE may be involved in stress response. Yeast also contains a member of

the family Swiss:P38254. Swiss:P30887 can complement a mutation in acid phosphatase, suggesting that members of this family could be phosphatases. Number of members: 17.

[1]Medline: 95014035. A new gene involved in stationary-phase survival located at 59 minutes on the Escherichia coli chromosome. Li C, Ichikawa JK, Ravetto JJ, Kuo HC, Fu JC, Clarke S; J Bacteriol 1994;176:6015-6022.

[2]Medline: 93046805. Complementation of Saccharomyces cerevisiae acid phosphatase mutation by a genomic sequence from the yeast Yarrowia lipolytica identifies a new phosphatase. Treton BY, Le Dall MT, Gaillardin CM; Curr Genet 1992;22:345-355.

1018. Synuclein - Synuclein

There are three types of synucleins in humans, these are called alpha, beta and gamma. Alpha synuclein has been found mutated in families with autosomal dominant Parkinson's disease. A peptide of alpha synuclein has also been found in amyloid plaques in Alzheimer's patients. Number of members: 12.

[1]Medline: 98424410. The synuclein family. Lavedan C; Genome Res 1998;8:871-880.

1019. (T-box) T-box domain signatures

PROSITE: PDOC00972. PROSITE cross-reference(s) PS01283; TBOX_1 PS01264; TBOX_2

A number of eukaryotic DNA-binding proteins contain a domain of about 170 to 190 amino acids known as the T-box domain [1,2,3] and which probably binds DNA. The T-box has first been found in the mice T locus (Brachyury) protein, a transcription factor involved in mesoderm differentiation. It has since been found in the following proteins:

-Vertebrate and invertebrate homologs of the T protein.

-Mammalian proteins TBX1 to TBX6.

-Mammalian protein TBR1 which is expressed specifically in brain.

-Xenopus laevis eomesodermin (eomes).

-Xenopus laevis Vegt (or Antipodean), a transcription factor that activates the expression of wnt-8, eomes and Brachyury.

-Chicken TbxT.

-Drosophila protein optomotor-blind (omb).

-Drosophila protein brachyenteron (byn) (also known as Trg), which is

required for the specification of the hindgut and anal pads.

-Drosophila protein H15.

-Caenorhabditis elegans protein tbx-12.

-Caenorhabditis elegans hypothetical proteins F21H11.3, F40H6.4, T07C4.2, T07C4.6 and
5   ZK177.10.

Two conserved regions were selected as signature patterns for the T-domain. The first region
corresponds to the N-terminal of the domain and the second one to the central part.

Consensus patternL-W-x(2)-[FC]-x(3,4)-[NT]-E-M-[LIV](2)-T-x(2)-G-[RG]-[KRQ]

10   Sequences known to belong to this class detected by the patternALL, except for C.elegans
ZK177.10.

Consensus pattern[LIVMYW]-H-[PADH]-[DEN]-[GS]-x(3)-G-x(2)-W-M-x(3)-[IVA]-x- F

Sequences known to belong to this class detected by the patternALL, except for C.elegans
tbx-12, ZK177.10 and Drosophila H15.

15

[1]Bollag R.J., Siegfried Z., Cebra-Thomas J.A., Garvey N., Davison E.M., Silver L.M. Nat.
Genet. 7:383-389(1994).

[2] Agulnik S.I., Garvey N., Hancock S., Ruvinsky I., Chapman D.L., Agulnik I., Bollag R.J.,
Papaioannou V.E., Silver L.M. Genetics 144:249-254(1996).

20   [3]Papaioannou V.E. Trends Genet. 13:212-213(1997).

1020. Toprim - Toprim domain

This is a conserved region from DNA primase. This corresponds to the Toprim domain
common to DnaG primases, topoisomerases, OLD family nucleases and RecR proteins [1].
25   Both DnaG motifs IV and V are present in the alignment, the DxD (V) motif may be involved
in Mg2+ binding and mutations to the conserved glutamate (IV) completely abolish DnaG
type primase activity [1]. DNA primase EC:2.7.7.6 is a nucleotidyltransferase it synthesizes
the oligoribonucleotide primers required for DNA replication on the lagging strand of the
replication fork; it can also prime the leading stand and has been implicated in cell division
30   [2]. Number of members: 133.

[1]Medline: 98391745. Toprim--a conserved catalytic domain in type IA and II
topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. Aravind L,
Leipe DD, Koonin EV; Nucleic Acids Res 1998;26:4205-4213.

[2]Medline: 97368180. Cloning and analysis of the dnaG gene encoding Pseudomonas putida DNA primase. Szafranski P, Smith CL, Cantor CR; Biochim Biophys Acta 1997;1352:243-248.

[3]Medline: 94124015. The Haemophilus influenzae dnaG sequence and conserved bacterial primase motifs. Versalovic J, Lupski JR; Gene 1993;136:281-286.

1021. TraB - TraB family

pAD1 is a hemolysin/bacteriocin plasmid originally identified in Enterococcus faecalis DS16. It encodes a mating response to a peptide sex pheromone, cAD1, secreted by recipient bacteria. Once the plasmid pAD1 is acquired, production of the pheromone ceases--a trait related in part to a determinant designated traB. However a related protein is found in C. elegans Swiss:Q94217, suggesting that members of the TraB family have some more general function. Number of members: 12.

[1]Medline: 94302142. Characterization of the determinant (traB) encoding sex pheromone shutdown by the hemolysin/bacteriocin plasmid pAD1 in Enterococcus faecalis. An FY, Clewell DB; Plasmid 1994;31:215-221.

1022. (Transpo_mutator) Transposases, Mutator family, signature

PROSITE: PDOC00770. PROSITE cross-reference(s) PS01007;
TRANSPOSASE_MUTATOR

Autonomous mobile genetic elements such as transposon or insertion sequences (IS) encode an enzyme, called transposase, required for excising and inserting the mobile element. On the basis of sequence similarities, transposases can be grouped into various families. One of these families has been shown [1,2,3,E1] to consist of transposases from the following elements:

-Mutator from Maize.

-Is1201 from Lactobacillus helveticus.

-Is905 from Lactococcus lactis.

-Is1081 from Mycobacterium bovis.

-Is6120 from Mycobacterium smegmatis.

-Is406 from Pseudomonas cepacia.

-IsRm3 from Rhizobium meliloti.

-IsRm5 from Rhizobium meliloti.

-Is256 from Staphylococcus aureus.

-IsT2 from Thiobacillus ferrooxidans.

The maize Mutator transposase (MudrA) is a protein of 823 amino acids; the bacterial transposases listed above are proteins of 300 to 420 amino acids. These proteins contain a conserved domain of about 130 residues; a signature pattern was derived from the most conserved part of this domain.

Consensus patternD-x(3)-G-[LIVMF]-x(6)-[STAV]-[LIVMFYW]-[PT]-x-[STAV]-x(2)-[QR]-x-C-x(2)-H. Sequences known to belong to this class detected by the patternALL.

[1]Eisen J.A., Benito M.-I., Walbot V. Nucleic Acids Res. 22:2634-2636(1994).

[2]Guilhot C., Gicquel B., Davies J., Martin C. Mol. Microbiol. 6:107-113(1992).

[3]Wood M.S., Byrne A., Lessie T.G. Gene 105:101-105(1991).

1023. Transposase_8 - Transposase

Transposase proteins are necessary for efficient DNA transposition. This family consists of various E. coli insertion elements and other bacterial transposases some of which are members of the IS3 family. Number of members: 58.

[1]Medline: 97324595. Genetic organization and transposition properties of IS511. D. A. Mullin, D. L. Zies, A. H. Mullin, N. Caballera & B. Ely; Mol Gen Genet 1997;254:456-463. [2]Medline: 97128810. The use of an improved transposon mutagenesis system for DNA sequencing leads to the characterization of a new insertion sequence of Streptomyces lividans 66. J. Fischer, H. Maier, P. Viell & J. Altenbuchner; Gene 1996;180:81-89.

[3]Medline: 97074647. Identification and nucleotide sequence of Rhizobium meliloti insertion sequence ISRm6, a small transposable element that belongs to the IS3 family. S. Zekri & N. Toro; Gene 1996;175:43-48.

1024. tRNA_int_endo - tRNA intron endonuclease

Members of this family cleave pre tRNA at the 5' and 3' splice sites to release the intron EC:3.1.27.9. Number of members: 8.

791

[1]Medline: 97344075. Properties of H. volcanii tRNA intron endonuclease reveal a relationship between the archaeal and eucaryal tRNA intron processing systems. Kleman-Leyer K, Armbruster DW, Daniels CJ; Cell 1997;89:839-847.

5   1025. Urease - Urease signatures

PROSITE: PDOC00133PROSITE cross-reference(s) PS01120; UREASE_1 PS00145; UREASE_2

Urease (EC 3.5.1.5) is a nickel-binding enzyme that catalyzes the hydrolysis of urea to carbon dioxide and ammonia [1]. Historically, it was the first enzyme to be crystallized (in

10   1926). It is mainly found in plant seeds, microorganisms and invertebrates. In plants, urease is a hexamer of identical chains. In bacteria [2], it consists of either two or three different subunits (alpha, beta and gamma).

Urease binds two nickel ions per subunit; four histidine, an aspartate and a carbamated-lysine serve as ligands to these metals; an additional histidine is involved in the

15   catalytic mechanism [3].

As signatures for this enzyme, a region that contains two histidine that bind one of the nickel ions and the region of the active site histidine was selected.

Consensus pattern T-[AY]-[GA]-[GAT]-[LIVM]-D-x-H-[LIVM]-H-x(3)-P [The two H's bind

20   nickel].Sequences known to belong to this class detected by the patternALL.

Consensus pattern[LIVM](2)-[CT]-H-[HN]-L-x(3)-[LIVM]-x(2)-D-[LIVM]-x-F-A [H is the active site residue]. Sequences known to belong to this class detected by the patternALL.

[1]Takishima K., Suga T., Mamiya G. Eur. J. Biochem. 175:151-165(1988).

25   [2]Mobley H.L.T., Husinger R.P. Microbiol. Rev. 53:85-108(1989).

[3]Jabri E., Carr M.B., Hausinger R.P., Karplus P.A. Science 268:998-1004(1995).

1026. Urease_beta - Urease beta subunit.

This subunit is known as alpha in Heliobacter. Number of members: 35.

30

[1]Medline: 95273988. The crystal structure of urease from Klebsiella aerogenes. Jabri E, Carr MB, Hausinger RP, Karplus PA; Science 1995;268:998-1004.

1027. UvrD-helicase - UvrD/REP helicase

The Rep family helicases are composed of four structural domains. The Rep family function as dimers. REP helicases catalyse ATP dependent unwinding of double stranded DNA to single stranded DNA. Swiss:P23478, Swiss:P08394 have large insertions near to the carboxy-terminus relative to other members of the family. Number of members: 52.

[1] Medline: 97433075. Major domain swiveling revealed by the crystal structures of complexes of E. coli Rep helicase bound to single-stranded DNA and ADP. Korolev S, Hsieh J, Gauss GH, Lohman TM, Waksman G; Cell 1997;90:635-647.

1028. V-type ATPase 116kDa subunit family (V_ATPase_sub_a)

This family consists of the 116kDa V-type ATPase (vacuolar (H+)-ATPases) subunits, as well as V-type ATP synthase subunit i. The V-type ATPases family are proton pumps that acidify intracellular compartments in eukaryotic cells for example yeast central vacuoles, clathrin-coated and synaptic vesicles. They have important roles in membrane trafficking processes [1]. The 116kDa subunit (subunit a) in the V-type ATPase is part of the V0 functional domain responsible for proton transport. The a subunit is a transmembrane glycoprotein with multiple putative transmembrane helices t has a hydrophilic amino terminal and a hydrophobic carboxy terminal [1,2]. It has roles in proton transport and assembly of the V-type ATPase complex [1,2]. This subunit is encoded by two homologous gene in yeast VPH1 and STV1 [2].
Number of members: 27

[1] Forgac M; Medline: 99240666 Structure and properties of the vacuolar (H+)-ATPases." J Biol Chem 1999;274:12951-12954.
[2] Forgac M; Medline: 99270697 Structure and properties of the clathrin-coated vesicle and yeast vacuolar V-ATPases." J Bioenerg Biomembr 1999;31:57-65.

1029. Viral (Superfamily 1) RNA helicase (Viral_helicase1)
Number of members: 260

[1] Koonin EV, Dolja VV; Medline: 94094568 Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences." Crit Rev Biochem Mol Biol 1993;28:375-430.

1030. Vesicular monoamine transporter (VMAT)

5 This family consists of various vesicular amine transporters with 12 transmembrane helices. These included vesicular acetylcholine transporters (VAChT) [3], and vesicular monoamine transporters (VMATs) [1,2] isoforms 1 adrenal and 2 brain (VMAT1 and VMAT2).

These proteins transport biogenic amines into synaptic vesicles or chromaffin granules [4]. VMATs pack monoamine neurotransmitters into secretary vesicles for regulated exocytotic

10 release, they also protect against the parkinsonian neurotoxins MPP+ by transporting it into vesicles preventing it from acting on mitochondria [1].

Also in the family is C. elegans UNC-17 a putative vesicular acetylcholine transporter mutations in UNC-17 cause impaired neuromuscular function, giving rise to jerky or

15 uncoordinated movement, [4].
Number of members: 15

[1] Krantz DE, Peter D, Liu Y, Edwards RH; Medline: 97197857 Phosphorylation of a vesicular monoamine transporter by casein kinase II." J Biol Chem 1997;272:6752-6759.

20 [2] Erickson JD, Varoqui H, Schafer MK, Modi W, Diebler MF, Weihe E, Rand J, Eiden LE, Bonner TI, Usdin TB; Medline: 94350930 Functional identification of a vesicular acetylcholine transporter and its expression from a 'cholinergic' gene locus." J Biol Chem 1994;269:21929-21932.

[3] Erickson JD, Schafer MK, Bonner TI, Eiden LE, Weihe E; Medline: 96209876 Distinct

25 pharmacological properties and distribution in neurons and endocrine cells of two isoforms of the human vesicular monoamine transporter." Proc Natl Acad Sci U S A 1996;93:5166-5171.
[4] Alfonso A, Grundahl K, Duerr JS, Han HP, Rand JB; Medline: 3342494 The Caenorhabditis elegans unc-17 gene: a putative vesicular acetylcholine transporter." Science 1993;261:617-619.

30

1031. WW/rsp5/WWP domain signature and profile. Cross-reference(s): PS01159; WW_DOMAIN_1; PS50020; WW_DOMAIN_2

The WW domain [1-4,E1] (also known as rsp5 or WWP) has been originally discovered as a short conserved region in a number of unrelated proteins, among them dystrophin, the gene responsible for Duchenne muscular dystrophy. The domain, which spans about 35 residues, is repeated up to 4 times in some proteins. It has been shown [5] to bind proteins with

5 particular proline-motifs, [AP]-P-P-[AP]-Y, and thus resembles somewhat SH3 domains. It appears to contain beta-strands grouped around four conserved aromatic positions; generally Trp. The name WW or WWP derives from the presence of these Trp as well as that of a conserved Pro. It is frequently associated with other domains typical for proteins in signal transduction processes.

10

Proteins containing the WW domain are listed below.

--Dystrophin, a multidomain cytoskeletal protein. Its longest alternatively spliced form consists of an N-terminal actin-binding domain, followed by 24 spectrin-like repeats, a

15 cysteine-rich calcium-binding domain and a C-terminal globular domain. Dystrophin form tetramers and is thought to have multiple functions including involvement in membrane stability, transduction of contractile forces to the extracellular environment and organization of membrane specialization. Mutations in the dystrophin gene lead to muscular dystrophy of Duchenne or Becker type. Dystrophin contains one WW domain C-terminal of the spectrin-

20 repeats.

--Utrophin, a dystrophin-like protein of unknown function.

--Vertebrate YAP protein is a substrate of an unknown serine kinase. It binds to the SH3 domain of the Yes oncoprotein via a proline-rich region. This protein appears in alternatively spliced isoforms, containing either one or two WW domains [6].

25 --Mouse NEDD-4 plays a role in the embryonic development and differentiation of the central nervous system. It contains 3 WW modules followed by a HECT domain. The human ortholog contains 4 WW domains, but the third WW domain is probably spliced resulting in an alternate NEDD-4 protein with only 3 WW modules [3].

--Yeast RSP5 is similar to NEDD-4 in its molecular organization. It contains an N-terminal

30 C2 domain (see <PDOC00380>), followed by a histidine-rich region, 3 WW domains and a HECT domain.

--Rat FE65, a transcription-factor activator expressed preferentially in liver. The activator domain is located within the N-terminal 232 residues of FE65, which also contain the WW domain.

--Yeast ESS1/PTF1, a putative peptidyl prolyl cis-trans isomerase from family ppiC (see <PDOC00840>). A related protein, dodo (gene dod) exists in Drosophila and in mammals (gene PIN1).

--Tobacco DB10 protein. The WW domain is located N-terminal to the region with

5   similarity to ATP-dependent RNA helicases.

--IQGAP, a human GTPase activating protein acting on ras. It contains an N-terminal domain similar to fly muscle mp20 protein and a C-terminal ras GTPase activator domain.

--Yeast pre-mRNA processing protein PRP40, Caenorhabditis elegans ZK1098.1 and fission yeast SpAC13C5.02 are related proteins with similarity to MYO2-type myosin, each

10   containing two WW-domains at the N-terminus.

--Caenorhabditis elegans hypothetical protein C38D4.5, which contains one WW module, a PH domain (see <PDOC50003>) and a C-terminal phosphatidylinositol 3-kinase domain.

--Yeast hypothetical protein YFL010c.

15   For the sensitive detection of WW domains, a profile was developed which spans the whole homology region as well as a pattern.

Description of pattern(s) and/or profile(s):

20   Consensus patternW-x(9,11)-[VFY]-[FYW]-x(6,7)-[GSTNE]-[GSTQCR]-[FYW]-x(2)-P.

[1] Bork P., Sudol M. Trends Biochem. Sci. 19:531-533(1994).

[2] Andre B., Springael J.Y. Biochem. Biophys. Res. Commun. 205:1201-1205(1994).

[3] Hofmann K.O., Bucher P. FEBS Lett. 358:153-157(1995).

25   [4] Sudol M., Chen H.I., Bougeret C., Einbond A., Bork P. FEBS Lett. 369:67-71(1995).

[5] Chen H.I., Sudol M. Proc. Natl. Acad. Sci. U.S.A. 92:7819-7823(1995).

[6] Sudol M., Bork P., Einbond A., Kastury K., Druck T., Negrini M., Huebner K., Lehman D. J. Biol. Chem. 270:14733-14741(1995).

30   1032. XPA protein signatures. cross-reference(s): XPA_1 PROSITE PS00752; PS00753;XPA_2.

Xeroderma pigmentosum (XP) [1] is a human autosomal recessive disease, characterized by a high incidence of sunlight-induced skin cancer. People's skin cells with this condition are hypersensitive to ultraviolet light, due

to defects in the incision step of DNA excision repair. There are a minimum of seven genetic complementation groups involved in this pathway: XP-A to XP-G. XP-A is the most severe form of the disease and is due to defects in a 30 Kd nuclear protein called XPA (or XPAC) [2].

5

The sequence of the XPA protein is conserved from higher eukaryotes [3] to yeast (gene RAD14) [4]. XPA is a hydrophilic protein of 247 to 296 amino-acid residues which has a C4-type zinc finger motif in its central section.

10   Two signature were developed patterns for XPA proteins. The first corresponds to the zinc finger region, the second to a highly conserved region located some 12 residues after the zinc finger region.

Consensus patternC-x-[DE]-C-x(3)-[LIVMF]-x(1,2)-D-x(2)-L-x(3)-F-x(4)-C-x(2)-C

15 Consensus pattern[LIVM](2)-T-[KR]-T-E-x-K-x-[DE]-Y-[LIVMF](2)-x-D-x-[DE]

[1] Tanaka K., Wood R.D. Trends Biochem. Sci. 19:83-86(1994).

[2] Miura N., Miyamoto I., Asahina H., Satokata I., Tanaka K., Okada Y. J. Biol. Chem. 266:19786-19789(1991).

20 [3] Shimamoto T., Kohno K., Tanaka K., Okada Y. Biochem. Biophys. Res. Commun. 181:1231-1237(1991).

[4] Bankmann M., Prakash L., Prakash S. Nature 355:555-558(1992).

1033. YCF9

25   This family consists of the hypothetical protein product of the YCF9 gene from chloroplasts and cyanobacteria. Number of members: 16

1034. (DUF15)

30 It is highly conserved between eubacteria and eukaryotes.

Number of members: 30

1035. Lumenal portion of Cytochrome b559, alpha (gene psbE) subunit. (cytochr_b559a)

This family is the lumenal portion of cytochrome b559 alpha chain, matches to this family should be accompanied by a match to the cytochr_b559 family also. The Prosite pattern pattern matches the transmembrane region of the cytochrome b559 alpha and beta subunits.

5 Number of members: 16

A. Asparaginase 2

10

Asparaginase II (L-asparagine aminohydrolase II) is an extracellular protein that may be associated with the cell wall and whose expression is affected by the availability of nitrogen. Asparaginase II catalyzes the reaction of L-Asparagine + H_2O = L-Aspartate + NH_3. As many leukemias have high requirements for aspartic acid, asparaginase II proteins are useful

15 as reagents for screening compounds for activity as leukemia chemotherapy products. Asparaginase II protein can also be over- or under-expressed to alter amino acid content in plant tissues or to modify nitrogen fixation and/or nitrogen metabolism in plants.

Ref: Bon et al. (1997) Appl Biochem Biotechnol 63-65: 203-12

20

B. Chloroa b-bind

Chlorophyll a-b binding proteins are located in the thylakoid membranes of the chloroplast and bind chlorophyll a and chlorophyll b, thereby triggering a chemical reaction

25 (photosynthesis). These proteins are useful in controlling the rate, efficiency and/or output of photosynthesis. Overexpression of chlorophyll a-b binding proteins is expected to increase the rate of photosynthesis.

Ref: Leutwiler et al. (1986) Nucleic Acids Res 14: 4051-64

30   Brandt et al. (1992) Plant Mol Biol 19: 699-703

C. DMRL synthase

DMRL Synthase (6,7-Dimethyl-8-Ribityllumazine Synthase) catalyzes the last step in riboflavin (Vitamin B_2) synthesis, condensing 5-amino-6-(1'-D)-ribityl-amino-2,4(1H, 3H)-Pyrimidinedione with L-3,4-Dihydroxy-2-Butanone 4-Phosphate producing 6,7-Dimethyl-8-(1-D-Ribityl)Luminazine . The enzyme forms a homopentamer. Engineering of these

5   proteins or those with homologous sequences/structures may allow control of the amounts of vitamin B_2 available in plants and/or accumulation of pigment, as well as altering reactions requiring hydrogen ion carriers/transmitters.

Ref: Garcia-Ramirez et al. (1995) J Biol Chem **270**: 23801-7

10

D. E1_N

These proteins are ATP-dependent DNA helicases that are required for initiation of viral DNA replication. They form a complex with the viral E2 protein. The E1-E2 complex binds

15   to the replication origin that contains binding sites for both proteins. The majority of sequences known for this group of proteins are from various papillomaviruses, a type of double stranded DNA virus. In plants, the prototype double stranded DNA virus is Cauliflower Mosaic virus (CaMV). Manipulation of these proteins, especially to produce variant proteins that form non-productive complexes, enables production of plants that are

20   resistant to infection by double stranded DNA viruses.

Ref: Yang et al. (1993) PNAS USA **90**: 5086-90

Ustav and Stenlund (1991) EMBO J **10**: 449-57

Callaway et al. (1996) Mol Plant Microbe Interact 9: 810-8

25

E. EF1_G

Elongation Factor-1 is composed of four subunits: alpha, beta, delta and gamma. Gamma subunits are presumed to play a role in anchoring the complex to other cellular components.

30   Studies of EF-1 genes in plants suggests that different forms of the EF-1 subunits may be expressed in particular organs or in response to stress. Manipulation of the activity of these proteins, either by altered expression level or by structural mutation, may result in the accumulation of a particular protein in a chosen organ or allow production of particular proteins during stress conditions.

799

Ref: Kinzy et al. (1994) NAR 22: 2703-7

Dunn et al. (1993) Plant Mol Biol 23: 221-5

Aguilar et al. (1991) Plant Mol Biol 17: 351-60

5

F.   ENV_polyprotein

This family comprises the envelope or coat proteins known from a number of different retroviruses. In mammalian species, retroviruses are responsible for diseases such as

10   leukemia and HIV. In plants, retroviruses are known in both monocot (e.g. Zeon-1) and dicot (e.g. Arabidopsis and tobacco) species and have been shown to induce mutant alleles at new loci. Engineering of plant ENV proteins may allow mobilization or targeting of endogenous or introduced retroviruses, in essence generating a new method for mutant production, gene tagging and the like.

15

Ref: Mamoun et al (1990) J Virol 64: 4180-8

Grandbastien et al. (1989) Nature 337: 376-80

Wright and Voytas (1998) Genetics 149: 703-15

20

G. Glycosyl_hydr9

Proteins having this domain (previously known as the glycosyl hydrolase family 5 domain) catalyze the endohydrolysis of 1,4-β-D-glucosidic linkages in cellulose. Numerous plant

25   proteins with this domain exist and are expressed in an organ specific manner. They are involved in the fruit ripening process, in cell elongation and plant reproduction. Modulation of the activity of these proteins, either by over- or under-expression or by mutation of the polypeptide, could be used to affect post-harvest physiology (e.g. rate of ripening) or for engineering reproductive sterility.

30

Ref: Giorda et al. (1990) Biochemistry 29: 7264-9

Tucker et al. (1988) Plant Physiol 88: 1257-62

Shani et al. (1997) 43: 837-42

Milligan and Gasser (1995) Plant Mol Biol 28: 691-711

H. Glycosyl_hydr14

5   The β-amylases (family 14 of glycosyl hydrolases) catalyze the hydrolysis of 1,4-α-glucosidic linkages in polysaccharides and remove successive maltose units from the non-reducing ends of the chains. Mutants of β-amylase in Arabidopsis exhibited altered degradation of starch throughout the diurnal cycle. In addition, the mutant phenotypes indicated that these enzymes not only affect carbohydrate metabolism/catabolism, but also

10   influence the amount of pigment stored within particular cells. Manipulation of the β-amylase genes enables control of plant pigmentation (for example, fibre pigment in cotton) as well as carbohydrate synthesis and degradation.

Ref: Zeeman et al. (1998) Plant J 15: 357-65

15   Hirano and Nakamura (1997) Plant Physiol 114: 5675-82

          Kitamoto et al. (1988) J Bacteriol 170: 5848-54

I.   Glycosyl_hydr15

20

Glycosyl hydrolases from family 15 (such as 1,4-Alpha-D-Glucan glucohydrolase,) catalyze the hydrolysis of terminal 1,4-linked alpha-D-glucose residues successively from the non-reducing ends of the chains resulting in the release of β–D-Glucose. In plants these proteins have been tied to the mobilization of the xyloglucan stored in the cotyledonary cell walls.

25   Proteins such as these could be varied to affect the rate of plant growth (for example during germination), storage and/or use of glucose and other sugars by plant tissues and alteration of the properties, such as elasticity, of plant cell walls.

Ref: Crombie et al. (1998) Plant J 15: 27-38

30   Hata et al. (1991) Agric Biol Chem 55: 941-9

J.   Glycosyl_hydr20

Members of the family 20 glycosyl hydrolases catalyze the hydrolysis of terminal non-reducing N-acetly-D-hexosamine residues in N-acetyl-β-D-hexosaminides. N-acetyl-β -glucosaminidase belongs to this family and exists in several different forms (consisting of various combinations of alpha and beta chains) depending on the organism. Family 20

5 glycosyl hydrolases have been implicated in lysosomal storage diseases (such as Sandhoff disease) and glycogen' storage disease in humans. These types of proteins are also responsible for the hydrolysis of chitin. In plants, these proteins could be useful in controlling carbohydrate catabolism, thereby influencing the amount of sugars available for storage and/or use in other metabolic pathways. In addition, it is possible that such proteins

10 could be used to engineer an endogenous insect protection mechanism, e.g. by secretion of a chitin-hydrolyzing composition by the plant.

Ref: Graham et al (1988) J Biol Chem 263: 16823-9

O'Dowd et al. (1988) Biochemistry 27: 5216-26

15

K. HMG box

The HMG box is a novel type of DNA-binding domain found in a diverse group of proteins. Numerous plant proteins contain this domain, such as the HMG1/2-like proteins. The

20 expression of some of these HMG proteins appears to be regulated by circadian rhythms and in a light dependent manner, occurring at higher levels in roots, for example and lower levels in light-grown tissues such as cotyledons. Generally, HMG proteins are thought to influence transcription regulation. In plants, HMGs are believed to have a role in maintaining patterns of circadian-regulated expression for other genes, suggesting that these proteins could be

25 exploited to control growth and development.

Ref: Laudet et al. (1993) Nucleic Acids Res 21: 2493-501

Zheng et al. (1993) Plant Mol Biol 23: 813-23

Grasser et al. (1993) Plant Mol Biol 23: 619-25

30

L. IL2

Interleukin-2 (IL-2)is produced in mammals by T cells in response to antigenic or mitogenic stimulation and is crucial for proper regulation and functioning of the immune response. IL-2 is capable of stimulating B cells, monocytes, lymphokine-activated killer cells, natural killer cells and glioma cells. Plant extracts have also been shown to stimulate the immune system

5 (for example, mistletoe therapy for human cancer). It is known that IL-2 is involved in feedback inhibition pathways that impact the inflammatory response as well as the growth inhibition of tumor reactive T cells. Plant proteins containing IL-2-like sequences are useful as immunity-based therapeutics, acting in a manner similar to IL-2 in mammals.

10 Ref: Heike et al. (1997) Scand J Immunol 45: 221-6

Ariel et al. (1998) J Immunol 161: 2465-72

Schink (1997) Anticancer Drugs 8 Suppl 1: S47-51

M. Oxidored_FMN

15

NADPH dehydrogenases catalyze the reaction NADPH + acceptor = NADP(+) + reduced acceptor. One member of this family is yeast old yellow enzyme" (OYE) and is thought to be involved in oxylipin metabolism. A second yeast family member is a protein that binds estrogen binding protein (EBP) in addition to exhibiting oxidoreductase activity. An

20 Arabidopsis homolog to OYE has been described and estrogen binding proteins in plants have been reported. Plant proteins from this class have the potential to be used to modify lipid metabolism/catabolism. These proteins may also have use as therapeutics for breast and prostate cancer, and other abnormal growth in steroid-sensitive tissues.

25 Ref: Baker et al. (1998) Proc Soc Exp Biol Med 217: 317-21

Schaller and Weiler (1997) J Biol Chem 272: 28066-72

Mandani et al. (1994) PNAS USA 91: 922-6

N. Oxidored_q2

30

The NADH-plastoquinone oxidoreductases catalyze the reaction NADH + plastoquinone = NAD(+) + plastoquinol. In plants these reactions occur in the chloroplast and are believed to participate in a chloroplast respiratory system. Here, the NDH complex is postulated to act as

a valve to remove excess reduction equivalents in the chloroplasts. Manipulation of these proteins may improve the rate or efficiency of photosynthesis.

Ref: Burrows et al. (1998) EMBO J 17: 868-76

5   Kofer et al (1998) Mol Gen Genet 258: 166-73

        Maier et al. (1995) J Mol Biol 251: 614-28

O. PABP

10   Polyadenylate binding proteins bind the poly (A) tail of mRNA. Plants, as exemplified by Arabidopsis, contain numerous PABP genes that are expressed in an organ-specific manner. For example, PABP2 is functional in roots and shoots, while PABP5 is expressed predominantly in immature flowers. The PABP proteins are implicated in numerous aspects of posttranscriptional regulation including mRNA turnover and translational initiation.

15   Control of activity of PABP proteins provides the ability to control the expression of various genes in particular organs during development.

Ref: Hilson et al (1993) Plant Physiol 103: 525-33

       Belostotsky and Meagher (1993) PNAS USA 90: 6686-90

20

P. Parvo coat

Parvoviruses are linear single-stranded DNA viruses that are encapsulated by three capsid proteins. Plants are susceptible to infection by single stranded DNA viruses such as Maize

25   streak virus (MSV) and various Gemini viruses. The coat proteins in these plant viruses are critical to the virus life cycle within the plant. For example, the coat protein of MSV is thought to be involved in intra- and inter-cellular movement within the plant. Engineering of proteins having similarity to parvoviral coat proteins, especially to produce proteins that interfere with maturation of the virus particle, enables the production of plants having better

30   resistance to natural plant single-stranded DNA viruses.

Ref: Liu et al. (1997) J Gen Virol 78: 1265-70

       Rohde et al. (1990) Virology 176: 648-51

O. Pkinase C

Plant serine/threonine protein kinases possessing this domain are expressed in all tissues and are known to undergo serine-specific autophosphorylation and specifically phosphorylate two ribosomal proteins, P14 and P16. During development, these proteins predominate during high metabolic activity in growing buds, root tips, leaf margins and germinating seeds. They are thought to be involved in the control of plant growth and development. In addition, two genes encoding proteins from this family have been described that help plant cells adapt during cold or high salt stresses. Consequently, engineering Pkinase C proteins provides a way to control general growth/development of the plant as well as a means to provide endogenous protection against environmental stresses.

Ref: Zhang et al. (1994) J Biol Chem 269: 17586-92

Mizoguchi et al. (1995) FEBS Lett 358: 199-204

R. REV

The REV proteins act post-transcriptionally to relieve negative repression of GAG and ENV production in retroviruses such as Human Immounodeficiency Virus type I (HIV-1). Plants contain retrovirus-like viruses such as pararetroviruses and retrotransposons (i.e. transposons having long terminal repeats). Plant retrotransposons in particular have been used to create mutations at various loci, thereby permitting gene isolation, gene tagging and the like. Manipulation of plant REV proteins enables control of transposition frequencies of corresponding transposable elements and provides a new tool for genetic engineering of plants.

Ref: Sodroski et al. (1986) Nature 321: 412-7

Franchini et al. (1989) PNAS USA 86: 2433-7

Marquet et al. (1995) 77: 113-24

Grandbastien et al. (1989) Nature 337: 376-80

Wright and Voytas (1998) Genetics 149: 703-15

S. RuBisCo small

Ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCo) catalyzes the initial step in the C3 photosynthetic carbon reduction cycle, adding carbon dioxide to D-ribulose 1,5-bisphosphate to form two molecules of 3-phospho-D-glycerate. RuBisCo is comprised of two subunits, one large which is synthesized in the chloroplast, and one small which is

5 synthesized in the cytoplasm and then transported in to the chloroplast. The expression of the small subunit of RuBisCo is light regulated. Manipulation of these proteins could increase the efficiency of photosynthesis or allow alterations in developmental timing.

Ref: Giuliano et al. (1988) PNAS USA 85: 7089-93

10   Dedonder et al. (1993) Plant Physiol 101: 801-8

T. Sialyltransf

Members of the CMP-N-acetylneuraminate-β-galactosamide-α-2,3-sialyltransferase family

15 catalyze the following reaction:

CMP-N-acetylneuraminate + β-D-galactosyl-1,3-N-acetyl-α-D-galactosaminyl-R = CMP + α-N-acetylneraminyl-2,3-β-D-galactosyl-1,3-N-acetyl-alpha-D-galactosaminyl-R. These proteins are though to be responsible for the synthesis of the sequence neurac-α-2,3-gal-β-1,3-galnac- found on sugar chains)-linked to threonine or serine and also as a terminal

20 sequence on certain gangliosides in mammalian cells. In plants, glycosyltransferases in the Golgi apparatus synthesize cell wall polysaccharides and elaborate the complex glycans of glycoproteins. Engineering of plant sialyltransferases allows targeting of proteins to particular cellular locations or enables the making of changes in cell wall structure.

25 Ref: Wee et al. (1998) Plant Cell 10: 1759-68

Lee et al. (1994) J Biol Chem 269: 10028-33

Kitagawa and Paulson (1994) J Biol Chem 269: 1394-401

U. Signal

30

Many plant proteins in this family contain sequences similar to those found in both components of the prokaryotic family of signal transducers known as the two-component systems. This suggests that activation may require a transfer of a phosphate group between

the transmitter domain and the receiver domain. One family member in Arabidopsis appears to be involved in ethylene (a plant hormone) signal transduction. Other proteins in this family appear to be involved in the regulation of gene transcription under conditions of environmental stress. Signal proteins can be exploited to affect plant growth and development 5 and/or control plant responses to stress conditions such as cold, nutrient availability, etc.

Ref: Chang et al. (1993) Science 262: 539-44

Nagaya et al. (1993) Gene 131: 119-124

Gottfert et al. (1990) PNAS USA 87: 2680-4

10

V. vMSA

vMSA proteins are major surface antigens presenting on the envelope of various retroviruses. Surface antigens of retroviruses are often involved in tropism of the virus. 15 Plants contain retrovirus-like viruses such as pararetroviruses and retrotransposons (i.e. transposons having long terminal repeats). Plant retrotransposons in particular have been used to create mutants at various loci, thereby permitting gene isolation, gene tagging and the like. Manipulation of plant vMSA proteins enables control of tropism of plant retroviruses that might be used for genetic engineering tools, thus enabling targeting of the virus to 20 particular species and/or tissues of plants.

Ref: Okamoto et al. (1988) J Gen Virol 69: 2575-83

Grandbastien et al. (1989) Nature 337: 376-80

Wright and Voytas (1998) Genetics 149: 703-15

25

W. zf-CCCH

This family of proteins is defined by having two CX(8)CX(5)CX(3)H-type zinc finger domains. These proteins cover a broad range of functions. For example, the COP1 protein 30 acts as a repressor of photomorphogenesis in darkness; light stimuli abolish this suppressive action. In addition, COP1 protein can function as a negative transcriptional regulator capable of direct interaction with components of the G-protein signaling pathway. As a second example, a zf-CCCH protein identified in Arabidopsis appears to be involved in the resistance to DNA damage induced by UV light and chemical DNA-damaging agents.

Overexpression of this class of proteins permits production of plants that are better suited to adverse environments. Manipulation of expression of zf-CCCH proteins functioning as transcriptional regulators, such as COP1, enables manipulation of some signal transduction pathways.

Ref: Pang et al. (1993) Nucleic Acids Res 21: 1647-53

Deng et al. (1992) Cell 71: 791-801

X. zf-RanBP

Proteins falling within this category contain many X-X-F-G and X-F-X-F-G repeats, and may contain RANBP1-like or PPIase domains. Plant proteins having domains similar to these include PAS1 and GMSTI. PAS1 has been shown to have dramatic developmental affects that appear to be correlated with both cell division and cell wall elongation. GMSTI has high identity to the yeast STI stress-inducible gene and has been shown to be heat inducible. Proteins such as these may be useful for controlling growth and form of development.

Ref: Vittorioso et al. (1998) Mol Cell Biol 18: 3034-43

Hernandez Torres et al. (1995) 27: 1221-6

Y. Peptidase M48.

Proteins belonging to this peptidase family are metalloproteases that bind zinc as a cofactor and are located in the membranes of the endoplasmic reticulum. They function in NH_2-terminal proteolytic processing, as shown for the yeast STE24 gene product. This gene is required for the correct processing of α-factor, a yeast pheromone. Family M48 peptidases also appear to be required for some prenylation reactions, mediating COOH-terminal CAAX processing. Prenylation reactions are believed to be involved in the regulation of protein-protein and protein-membrane interactions. As an example, RAS GTPase activity is regulated in part by localization to the inner side of the plasma membrane upon prenylation. In plants, proteins from this family could be involved in pollen-stigma interactions such as those mediating self-pollenation vs. outcrossing, or could be members of several secondary metabolism pathways.

Ref: Fujimura-Kamada et al. (1997) J Cell Biol. 136: 271-85. Tam et al. (1998) J Cell Biol. 142: 635-49.

5   ### Z. DNA Pol Viral N

The DNA pol Viral N domain is located at the N-terminal region of DNA polymerase isolated from several retroid viruses such as the Cauliflower Mosaic Virus. The domain motif has also been found in numerous other species from humans to cyanobacteria. In these organisms, this motif seems to be associated with two types of sequences; retrotransposons

10   and mitochondrial genes. In the mitochondrial sequences this domain is potentially involved in the self-splicing conducted by group II introns. Various manipulations of this gene in plants allows control of the numerous retrotransposons endogenous to plant genomes or allows engineering of mitochondrial function, especially to increase efficiency of energy utilization by cells.

15

REF: Chapdelaine and Bonen (1991) Cell 65: 465-72

Ferat and Miche (1993) Nature 364: 358-61

Wilson et al. (1994) 368: 32-8

Cambareri et al. (1994) 242: 658-65

20   Gaardner et al. (1981) NAR 9: 2871-2888

Cummings et al. (1990) Curr Genet 17: 375-402

Hattori et al. (1986) Nature 321: 625-8

Aa.   Calpain_inhib

25   This domain is found in calpastatin, an inhibitor protein specific for calpain. Calpain is a non-lysosomal calcium-dependent intracellular protease that appears to be involved in the dynamic changes of the cytoskeleton, especially actin-related structures, during early *Drosophila* embryogenesis [1]. Calpastatins co-exist in cells with calpains and the subcellular distribution of calpastatin is thought to be important to calpain regulation [2]. In plants

30   calpains and calpastatins could be involved in embryogenesis and non-embryogenic organ reiteration. Mutations occurring in calpain inhibitor repeat domains would produce developmental abnormalities such as abnormal leaf, root or flower development.

Refs

1   Emori Y and Saigo K (1994) J Biol Chem 269: 25137-42.

2   Mellgren RL, Lane RD, Mericle MT (1989) Biochim Biophys Acta 999: 71-77.

Ab.   chorismate_bind

5   Chorismate binding domains are present in plant anthranilate synthase (AS) genes. AS genes catalyze the first step in the biosynthesis of tryptophan by converting chorismate and L-glutamine to anthranilate, pyruvate and L-glutamate. Some of these genes are involved in feedback inhibition by tryptophan [1] while some are feedback insensitive [2]. In Arabidopsis, two AS genes have overlapping, but different distributions. One of these AS

10   genes is induced by wounding and bacterial pathogen infiltration [1]. Mutations in the chorismate binding domain would affect the production of tryptophan and could influence the plant's defense system. AS gene products can be used for *in vitro* synthesis of tryptophan and tryptophan derivatives.

15   Refs

1   Niyogi KK, Fink GR (1992) Plant Cell 4: 721-33.

2   Song HS, Brotherton JE, Gonzales RA, Wilholm JM (1998) Plant Physiol 117:533-43.

20   ### Ac.   late_protein_L2

Papillomaviruses are encapsulated double stranded DNA viruses. Plants are susceptible to infection by double stranded DNA viruses such as Cauliflower Mosaic virus (CaMV). The coat proteins in these plant viruses are critical to the virus life cycle within the plant. For example, the coat protein of CaMV is thought to be involved in intra- and inter-cellular

25   movement within the plant [1]. Engineering of proteins having similarity to papillomavirus coat proteins may enable the production of plants having better resistance to natural plant double stranded DNA viruses.

Refs

30   1   Thompson SR, Melcher U (1993) J Gen Virol 74: 1141-8.

Ad.   Peptidase_M41

Proteins belonging to this peptidase family are metalloproteases that bind zinc as a cofactor and are integral membrane proteins. They seem to be involved in the degradation of carboxy-

terminal-tagged cytoplasmic proteins. In plants, these proteins are located in the thylakoid membranes of the chloroplasts, their expression is light regulated and they are thought to be involved in degradation of soluble stromal proteins and turn-over of thylkoid proteins [1]. Manipulation of expression and structure of these proteins would have effects on the

5   efficiency of photosynthesis and the development of chloroplasts.

Refs

1   Lindahl M, Tabak s, Cseke L, Pichersky E, Andersson B, Adam Z (1996) J Biol Chem 271: 29329-34.

10

Ae.   UPF0051

There is some evidence that, in plants, proteins in this family are involved in ATP synthesis in chloroplasts [1, 2]. Mutations in these proteins or altering their expression would affect the efficiency of photosynthesis and energy production.

15

Refs

1   Kostrzewa M, Zetsche K (1992) J Mol Biol 227: 961-70.

2   Kostrzewa M, Zetsche K (1993) Plant Mol Biol 23: 67-76

20   ### Af.   E7

Papillomaviruses are encapsulated double stranded DNA viruses. The Papillomavirus early protein 7 (E7) is known as a potent immortalizing and transforming agent. Transformation by E7 is thought to be mediated by the physical association of E7 with cellular proteins regulating entry into the cell cycle [1]. The result is entry into the cell cycle and suppression

25   of terminal differentiation in mammalian cells. Thus, engineering of proteins having similarity to papillomavirus E7 protein enables the production of plants having altered cellular proliferation characteristics and possibly altered morphology. For example, overexpression of E7-like proteins would be expected to result in proliferation of cells of the tissue in which the E7 protein is expressed, perhaps with suppression of differentiation

30   events. Thus, for example, overexpression of E7-like proteins in meristem cells can result in taller plants and suppression of leafing and/or flowering.

Refs

1   Zwerschke W, Jansen-Durr P Adv Cancer Res 2000;78:1-29

Ag.   Peptidase U7

This protein is known to be an integral membrane protein in the cyanobacterium Synechocystis where it functions to digest cleaved signal peptides [1]. This activity is necessary to maintain proper secretion of mature proteins across the membrane. In higher plants this protein may be present in the plastid or chloroplast membranes where it would function by enabling protein movement into and out of the chloroplasts. Mutations in this protein would be expected to affect the development of plastids, including chloroplasts, or alter the energy transfer system within the chloroplasts, thereby affecting growth and development.

Refs

1   Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S (1996) DNA Res 3:109-36.

Ah.   5'-3' Exonuclease

The 5'-3' exonuclease domain is one found in bacterial DNA polymerases I and in yeast DNA repair enzymes such as Exonuclease I. Yeast Exo I is involved in mitotic recombination and also includes a domain that interacts with the mismatch repair protein MSH2.   The 5'-3' exonuclease domain is also present in XPG DNA repair enzymes in humans and in yeast RAD9 protein. Defects in XPG proteins result in Xeroderma Pigmentosum. Thus defects in 5'-3' exonuclease domain-containing proteins in plants are expected to lead to defects in DNA repair and corresponding high spontaneous and inducible mutation rates. Consensus sequence:

IMKKKLLLVDGSSLAFRAFFALPPLTNSAGEPTNAVYGFLKMLIKLIEQEQPTHIAVV
FDAKAKTFRHELYEGYKAGRAP
TPDELREQIPLIKELLDALGIPLLEVAGYEADDVIGTLAKLAEKEGYEVLIVTGDRDLL
QLVSDHVTVIITKKGIAEFTL
FTPEAVIEKYGLTPEQIIDYKALMGDSSDNIPGVKGIGEKTAAKLLQEYGSLEGIYANL
DKLKGKKLREKLLAHKEDAKL
SRDLATIKTDVPLDLTLDDLRLPDPDRDALDLLFDE

Ref:

Fiorentini P. et al. RT. Mol. Cell. Biol. 17:2764-2773(1997).

Tishkoff et al. Cancer Res. 0:0-0(1998).

Macinnes M.A. et al. Mol. Cell. Biol. 13:6393-6402(1993).

AA. Activities of Polypeptides Comprising Signal Peptides

Polypeptides comprising signal peptides are a family of proteins that are typically targeted to (1) a particular organelle or intracellular compartment, (2) interact with a

5   particular molecule or (3) for secretion outside of a host cell. Example of polypeptides comprising signal peptides include, without limitation, secreted proteins, soluble proteins, receptors, proteins retained in the ER, etc.

These proteins comprising signal peptides are useful to modulate ligand-receptor

10   interactions, cell-to-cell communication, signal transduction, intracellular communication, and activities and/or chemical cascades that take part in an organism outside or within of any particular cell.

One class of such proteins are soluble proteins which are transported out of the cell.

15   These proteins can act as ligands that bind to receptor to trigger signal transduction or to permit communication between cells.

Another class is receptor proteins which also comprise a retention domain that lodges the receptor protein in the membrane when the cell transports the receptor to the surface of

20   the cell. Like the soluble ligands, receptors can also modulate signal transduction and communication between cells.

In addition the signal peptide itself can serve as a ligand for some receptors. An example is the interaction of the ER targeting signal peptide with the signal recognition

25   particle (SRP). Here, the SRP binds to the signal peptide, halting translation, and the resulting SRP complex then binds to docking proteins located on the surface of the ER, prompting transfer of the protein into the ER.

A description of signal peptide residue composition is described below in Subsection

30   IV.C.1.

III. Methods of Modulating Polypeptide Production

It is contemplated that polynucleotides of the invention can be incorporated into a host cell or in-vitro system to modulate polypeptide production. For instance, the SDFs prepared as described herein can be used to prepare expression cassettes useful in a number of
5 techniques for suppressing or enhancing expression.

An example are polynucleotides comprising sequences to be transcribed, such as coding sequences, of the present invention can be inserted into nucleic acid constructs to modulate polypeptide production. Typically, such sequences to be transcribed are heterologous to at least one element of the nucleic acid construct to generate a chimeric gene
10 or construct.

Another example of useful polynucleotides are nucleic acid molecules comprising regulatory sequences of the present invention. Chimeric genes or constructs can be generated when the regulatory sequences of the invention linked to heterologous sequences in a vector construct. Within the scope of invention are such chimeric gene and/or constructs.

15 Also within the scope of the invention are nucleic acid molecules, whereof at least a part or fragment of these DNA molecules are presented in TABLE 1 of the present application, and wherein the coding sequence is under the control of its own promoter and/or its own regulatory elements. Such molecules are useful for transforming the genome of a host cell or an organism regenerated from said host cell for modulating polypeptide production.

20 Additionally, a vector capable of producing the oligonucleotide can be inserted into the host cell to deliver the oligonucleotide.

More detailed description of components to be included in vector constructs are described both above and below.

Whether the chimeric vectors or native nucleic acids are utilized, such
25 polynucleotides can be incorporated into a host cell to modulate polypeptide production. Native genes and/or nucleic acid molecules can be effective when exogenous to the host cell.

Methods of modulating polypeptide expression includes, without limitation:

Suppression methods, such as

Antisense

30 Ribozymes

Co-suppression

Insertion of Sequences into the Gene to be Modulated

Regulatory Sequence Modulation.

as well as Methods for Enhancing Production, such as

Insertion of Exogenous Sequences; and

Regulatory Sequence Modulation.

5   III.A.   Suppression

Expression cassettes of the invention can be used to suppress expression of

endogenous genes which comprise the SDF sequence. Inhibiting expression can be useful,

for instance, to tailor the ripening characteristics of a fruit (Oeller et al., *Science* 254:437

(1991)) or to influence seed size_(WO98/07842) or to provoke cell ablation (Mariani et al.,

10   Nature 357: 384-387 (1992).

As described above, a number of methods can be used to inhibit gene expression in

plants, such as antisense, ribozyme, introduction of exogenous genes into a host cell,

insertion of a polynucleotide sequence into the coding sequence and/or the promoter of the

endogenous gene of interest, and the like.

15   III.A.1. Antisense

An expression cassette as described above can be transformed into host cell or

plant to produce an antisense strand of RNA. For plant cells, antisense RNA inhibits gene

expression by preventing the accumulation of mRNA which encodes the enzyme of interest, *see,*

e.g., Sheehy et al., *Proc. Nat. Acad. Sci.* USA, 85:8805 (1988), and Hiatt et al., U.S. Patent No.

20   4,801,340.

          III.A.2. Ribozymes

Similarly, ribozyme constructs can be transformed into a plant to cleave mRNA

and down-regulate translation.

          III.A.3. Co-Suppression

25   Another method of suppression is by introducing an exogenous copy of the gene

to be suppressed. Introduction of expression cassettes in which a nucleic acid is configured in

the sense orientation with respect to the promoter has been shown to prevent the accumulation of

mRNA. A detailed description of this method is described above.

          III.A.4.   Insertion of Sequences into the Gene to be Modulated

Yet another means of suppressing gene expression is to insert a polynucleotide into the gene of interest to disrupt transcription or translation of the gene.

Homologous recombination could be used to target a polynucleotide insert to a gene using the Cre-Lox system (A.C. Vergunst et al., *Nucleic Acids Res.* 26:2729 (1998), A.C.

5   Vergunst et al., *Plant Mol. Biol.* 38:393 (1998), H. Albert et al., *Plant J.* 7:649 (1995)).

In addition, random insertion of polynucleotides into a host cell genome can also be used to disrupt the gene of interest. Azpiroz-Leehan et al., *Trends in Genetics* 13:152 (1997). In this method, screening for clones from a library containing random insertions is preferred for identifying those that have polynucleotides inserted into the gene of interest. Such screening can

10   be performed using probes and/or primers described above based on sequences from TABLE 1, fragments thereof, and substantially similar sequence thereto. The screening can also be performed by selecting clones or any transgenic plants having a desired phenotype.

III.A.5. Regulatory SequenceModulation

The SDFs described in Table 1, and fragments thereof are examples of

15   nucleotides of the invention that contain regulatory sequences that can be used to suppress or inactivate transcription and/or translation from a gene of interest as discussed in I.C.5.

III.A.6. Genes Comprising Dominant-Negative Mutations

When suppression of production of the endogenous, native protein is desired it

20   is often helpful to express a gene comprising a dominant negative mutation. Production of protein variants produced from genes comprising dominant negative mutations is a useful tool for research Genes comprising dominant negative mutations can produce a variant polypeptide which is capable of competing with the native polypeptide, but which does not produce the native result. Consequently, over expression of genes comprising these mutations

25   can titrate out an undesired activity of the native protein. For example, The product from a gene comprising a dominant negative mutation of a receptor can be used to constitutively activate or suppress a signal transduction cascade, allowing examination of the phenotype and thus the trait(s) controlled by that receptor and pathway. Alternatively, the protein arising from the gene comprising a dominant-negative mutation can be an inactive enzyme still capable

30   of binding to the same substrate as the native protein and therefore competes with such native protein.

Products from genes comprising dominant-negative mutations can also act upon the native protein itself to prevent activity. For example, the native protein may be active only as a homo-multimer or as one subunit of a hetero-multimer. Incorporation of an inactive subunit into the multimer with native subunit(s) can inhibit activity.

5   Thus, gene function can be modulated in host cells of interest by insertion into these cells vector constructs comprising a gene comprising a dominant-negative mutation.

III.B.   Enhanced Expression

Enhanced expression of a gene of interest in a host cell can be accomplished by either (1) insertion of an exogenous gene; or (2) promoter modulation.

10   III.B.1. Insertion of an Exogenous Gene

Insertion of an expression construct encoding an exogenous gene can boost the number of gene copies expressed in a host cell.

Such expression constructs can comprise genes that either encode the native protein that is of interest or that encode a variant that exhibits enhanced activity as compared to

15 the native protein. Such genes encoding proteins of interest can be constructed from the sequences from TABLE 1, fragments thereof, and substantially similar sequence thereto.

Such an exogenous gene can include either a constitutive promoter permitting expression in any cell in a host organism or a promoter that directs transcription only in particular cells or times during a host cell life cycle or in response to environmental stimuli.

20   III.B.2. Regulatory Sequence Modulation

The SDFs of Table 1, and fragments thereof, contain regulatory sequences that can be used to enhance expression of a gene of interest. For example, some of these sequences contain useful enhancer elements. In some cases, duplication of enhancer elements or insertion of exogenous enhancer elements will increase expression of a desired gene from a particular

25 promoter. As other examples, all II promoters require binding of a regulatory protein to be activated, while some promoters may need a protein that signals a promoter binding protein to expose a polymerase binding site. In either case, over-production of such proteins can be used to enhance expression of a gene of interest by increasing the activation time of the promoter.

Such regulatory proteins are encoded by some of the sequences in TABLE 1,

30 fragments thereof, and substantially similar sequences thereto.

Coding sequences for these proteins can be constructed as described above.

IV. Gene Constructs and Vector Construction

To use isolated SDFs of the present invention or a combination of them or parts and/or mutants and/or fusions of said SDFs in the above techniques, recombinant DNA vectors which comprise said SDFs and are suitable for transformation of cells, such as plant cells, are usually prepared. The SDF construct can be made using standard recombinant DNA techniques (Sambrook et al. 1989) and can be introduced to the species of interest by *Agrobacterium*-mediated transformation or by other means of transformation (*e.g.*, particle gun bombardment) as referenced below.

The vector backbone can be any of those typical in the art such as plasmids, viruses, artificial chromosomes, BACs, YACs and PACs and vectors of the sort described by

(a) **BAC:** Shizuya et al., Proc. Natl. Acad. Sci. USA 89: 8794-8797 (1992); Hamilton et al., Proc. Natl. Acad. Sci. USA 93: 9975-9979 (1996);

(b) **YAC:** Burke et al., Science 236:806-812 (1987);.

(c) **PAC:** Sternberg N. et al., Proc Natl Acad Sci U S A. Jan;87(1):103-7 (1990);

(d) **Bacteria-Yeast Shuttle Vectors:** Bradshaw et al., Nucl Acids Res 23: 4850-4856 (1995);

(e) **Lambda Phage Vectors**: Replacement Vector, e.g., Frischauf et al., J. Mol Biol 170: 827-842 (1983); or Insertion vector, e.g., Huynh et al., In: Glover NM (ed) DNA Cloning: A practical Approach, Vol.1 Oxford: IRL Press (1985);

(f) **T-DNA gene fusion vectors** :Walden et al., Mol Cell Biol 1: 175-194 (1990); and

(g) **Plasmid vectors:** Sambrook et al., infra.

Typically, a vector will comprise the exogenous gene, which in its turn comprises an SDF of the present invention to be introduced into the genome of a host cell, and which gene may be an antisense construct, a ribozyme construct chimeraplast, or a coding sequence with any desired transcriptional and/or translational regulatory sequences, such as promoters, UTRs, and 3' end termination sequences. Vectors of the invention can also include origins of replication, scaffold attachment regions (SARs), markers, homologous sequences, introns, etc.

A DNA sequence coding for the desired polypeptide, for example a cDNA sequence encoding a full length protein, will preferably be combined with transcriptional and translational

initiation regulatory sequences which will direct the transcription of the sequence from the gene in the intended tissues of the transformed plant.

For example, for over-expression, a plant promoter fragment may be employed that will direct transcription of the gene in all tissues of a regenerated plant.   Alternatively, the plant

5   promoter may direct transcription of an SDF of the invention in a specific tissue (tissue-specific promoters) or may be otherwise under more precise environmental control (inducible promoters).

If proper polypeptide production is desired, a polyadenylation region at the 3'-end of the coding region is typically included. The polyadenylation region can be derived from the natural

10   gene, from a variety of other plant genes, or from T-DNA.

The vector comprising the sequences from genes or SDF or the invention may comprise a marker gene that confers a selectable phenotype on plant cells. The vector can include promoter and coding sequence, for instance. For example, the marker may encode biocide resistance, particularly antibiotic resistance, such as resistance to kanamycin, G418,

15   bleomycin, hygromycin, or herbicide resistance, such as resistance to chlorosulfuron or phosphinotricin.

IV.A.   Coding Sequences

Generally, the sequence in the transformation vector and to be introduced into the genome of the host cell does not need to be absolutely identical to an SDF of the present

20   invention. Also, it is not necessary for it to be full length, relative to either the primary transcription product or fully processed mRNA. Furthermore, the introduced sequence need not have the same intron or exon pattern as a native gene. Also, heterologous non-coding segments can be incorporated into the coding sequence without changing the desired amino acid sequence of the polypeptide to be produced.

25   IV.B.   Promoters

As explained above, introducing an exogenous SDF from the same species or an orthologous SDF from another species can modulate the expression of a native gene corresponding to that SDF of interest. Such an SDF construct can be under the control of either a constitutive promoter or a highly regulated inducible promoter (*e.g.*, a copper

30   inducible promoter). The promoter of interest can initially be either endogenous or heterologous to the species in question. When re-introduced into the genome of said species, such promoter becomes exogenous to said species. Over-expression of an SDF transgene can

lead to co-suppression of the homologous endogeneous sequence thereby creating some alterations in the phenotypes of the transformed species as demonstrated by similar analysis of the chalcone synthase gene (Napoli et al., *Plant Cell* 2:279 (1990) and van der Krol et al., *Plant Cell* 2:291 (1990)). If an SDF is found to encode a protein with desirable

5 characteristics, its over-production can be controlled so that its accumulation can be manipulated in an organ- or tissue-specific manner utilizing a promoter having such specificity.

Likewise, if the promoter of an SDF (or an SDF that includes a promoter) is found to be tissue-specific or developmentally regulated, such a promoter can be utilized to drive or

10 facilitate the transcription of a specific gene of interest (*e.g.*, seed storage protein or root-specific protein). Thus, the level of accumulation of a particular protein can be manipulated or its spatial localization in an organ- or tissue- specific manner can be altered.

IV. C Signal Peptides

15 SDFs of the present invention containing signal peptides are indicated in Table 1. In some cases it may be desirable for the protein encoded by an introduced exogenous or orthologous SDF to be targeted (1) to a particular organelle intracellular compartment, (2) to interact with a particular molecule such as a membrane molecule or (3) for secretion outside of the cell harboring the introduced SDF. This will be accomplished using a signal peptide.

20 Signal peptides direct protein targeting, are involved in ligand-receptor interactions and act in cell to cell communication. Many proteins, especially soluble proteins, contain a signal peptide that targets the protein to one of several different intracellular compartments. In plants, these compartments include, but are not limited to, the endoplasmic reticulum (ER), mitochondria, plastids (such as chloroplasts), the vacuole, the Golgi apparatus, protein

25 storage vessicles (PSV) and, in general, membranes. Some signal peptide sequences are conserved, such as the Asn-Pro-Ile-Arg amino acid motif found in the N-terminal propeptide signal that targets proteins to the vacuole (Marty (1999) *The Plant Cell* 11: 587-599). Other signal peptides do not have a consensus sequence *per se*, but are largely composed of hydrophobic amino acids, such as those signal peptides targeting proteins to the ER (Vitale

30 and Denecke (1999) *The Plant Cell* 11: 615-628). Still others do not appear to contain either a consensus sequence or an identified common secondary sequence, for instance the chloroplast stromal targeting signal peptides (Keegstra and Cline (1999) *The Plant Cell* 11: 557-570). Furthermore, some targeting peptides are bipartite, directing proteins first to an organelle and then to a membrane within the organelle (e.g. within the thylakoid lumen of the

chloroplast; see Keegstra and Cline (1999) *The Plant Cell* 11: 557-570). In addition to the diversity in sequence and secondary structure, placement of the signal peptide is also varied. Proteins destined for the vacuole, for example, have targeting signal peptides found at the N-terminus, at the C-terminus and at a surface location in mature, folded proteins. Signal

5   peptides also serve as ligands for some receptors.

These characteristics of signal proteins can be used to more tightly control the phenotypic expression of introduced SDFs. In particular, associating the appropriate signal sequence with a specific SDF can allow sequestering of the protein in specific organelles (plastids, as an example), secretion outside of the cell, targeting interaction with particular

10   receptors, etc. Hence, the inclusion of signal proteins in constructs involving the SDFs of the invention increases the range of manipulation of SDF phenotypic expression. The nucleotide sequence of the signal peptide can be isolated from characterized genes using common molecular biological techniques or can be synthesized in vitro.

In addition, the native signal peptide sequences, both amino acid and nucleotide, described in Table 1 can be used to modulate polypeptide transport. Further variants of the native signal peptides described in Table 1 are contemplated. Insertions, deletions, or substitutions can be made. Such variants will retain at least one of the functions of the native signal peptide as well as exhibiting some degree of sequence identity to the native sequence.

Also, fragments of the signal peptides of the invention are useful and can be fused with other signal peptides of interest to modulate transport of a polypeptide.

V.   Transformation Techniques

15   A wide range of techniques for inserting exogenous polynucleotides are known for a number of host cells, including, without limitation, bacterial, yeast, mammalian, insect and plant cells.

Techniques for transforming a wide variety of higher plant species are well known and described in the technical and scientific literature. *See, e.g.* Weising et al., *Ann. Rev. Genet.*

20   22:421 (1988); and Christou, Euphytica, v. 85, n.1-3:13-27, (1995).

DNA constructs of the invention may be introduced into the genome of the desired plant host by a variety of conventional techniques. For example, the DNA construct may be introduced directly into the genomic DNA of the plant cell using techniques such as electroporation and microinjection of plant cell protoplasts, or the DNA constructs can be

25   introduced directly to plant tissue using ballistic methods, such as DNA particle bombardment. Alternatively, the DNA constructs may be combined with suitable T-DNA flanking regions and

introduced into a conventional *Agrobacterium tumefaciens* host vector. The virulence functions of the *Agrobacterium tumefaciens* host will direct the insertion of the construct and adjacent marker into the plant cell DNA when the cell is infected by the bacteria (McCormac et al., *Mol. Biotechnol.* 8:199 (1997); Hamilton, *Gene* 200:107 (1997)); Salomon et al. *EMBO J.* 3:141

5   (1984); Herrera-Estrella et al. *EMBO J.* 2:987 (1983).

Microinjection techniques are known in the art and well described in the scientific and patent literature. The introduction of DNA constructs using polyethylene glycol precipitation is described in Paszkowski et al. *EMBO J.* 3:2717 (1984). Electroporation techniques are described in Fromm et al. *Proc. Natl Acad. Sci. USA* 82:5824 (1985). Ballistic transformation

10   techniques are described in Klein et al. *Nature* 327:773 (1987). *Agrobacterium tumefaciens*-mediated transformation techniques, including disarming and use of binary or co-integrate vectors, are well described in the scientific literature. See, for example Hamilton, *CM.*, *Gene* 200:107 (1997); Müller et al. *Mol. Gen. Genet.* 207:171 (1987); Komari et al. *Plant J.* 10:165 (1996); Venkateswarlu et al. *Biotechnology* 9:1103 (1991) and Gleave, *AP.*, *Plant Mol.*

15   *Biol.* 20:1203 (1992); Graves and Goldman, *Plant Mol. Biol.* 7:34 (1986) and Gould et al., *Plant Physiology* 95:426 (1991).

Transformed plant cells which are derived by any of the above transformation techniques can be cultured to regenerate a whole plant that possesses the transformed genotype and thus the desired phenotype such as seedlessness. Such regeneration techniques rely on

20   manipulation of certain phytohormones in a tissue culture growth medium, typically relying on a biocide and/or herbicide marker which has been introduced together with the desired nucleotide sequences. Plant regeneration from cultured protoplasts is described in Evans et al., *Protoplasts Isolation and Culture* in Handbook of Plant Cell Culture," pp. 124-176, MacMillan Publishing Company, New York, 1983; and Binding, *Regeneration of Plants, Plant Protoplasts,* pp. 21-73,

25   CRC Press, Boca Raton, 1988. Regeneration can also be obtained from plant callus, explants, organs, or parts thereof. Such regeneration techniques are described generally in Klee et al. *Ann. Rev. of Plant Phys.* 38:467 (1987). Regeneration of monocots (rice) is described by Hosoyama et al. (*Biosci. Biotechnol. Biochem.* 58:1500 (1994)) and by Ghosh et al. (*J. Biotechnol.* 32:1 (1994)). The nucleic acids of the invention can be used to confer desired traits on essentially any

30   plant.

Thus, the invention has use over a broad range of plants, including species from the genera *Anacardium, Arachis, Asparagus, Atropa, Avena, Brassica, Citrus, Citrullus, Capsicum, Carthamus, Cocos, Coffea, Cucumis, Cucurbita, Daucus, Elaeis, Fragaria, Glycine, Gossypium, Helianthus, Heterocallis, Hordeum, Hyoscyamus, Lactuca, Linum, Lolium,Lupinus,*

Lycopersicon, Malus, Manihot, Majorana, Medicago, Nicotiana, Olea, Oryza, Panieum,

Pannesetum, Persea, Phaseolus, Pistachia, Pisum, Pyrus, Prunus, Raphanus, Ricinus, Secale,

Senecio, Sinapis, Solanum, Sorghum, Theobromus, Trigonella, Triticum, Vicia, Vitis, Vigna,

and, Zea.

5   One of skill will recognize that after the expression cassette is stably incorporated in

transgenic plants and confirmed to be operable, it can be introduced into other plants by

sexual crossing. Any of a number of standard breeding techniques can be used, depending

upon the species to be crossed.

The particular sequences of SDFs identified are provided in the attached TABLE 1.

10   One of ordinary skill in the art, having this data, can obtain cloned DNA fragments, synthetic

DNA fragments or polypeptides constituting desired sequences by recombinant methodology

known in the art or described herein.

EXAMPLES

The invention is illustrated by way of the following examples. The invention is not

15   limited by these examples as the scope of the invention is defined solely by the claims

following.

EXAMPLE 1: cDNA PREPARATION

A number of the nucleotide sequences disclosed in TABLE 1 herein as representative of

the SDFs of the invention can be obtained by sequencing genomic DNA (gDNA) and/or cDNA

20   from corn plants grown from HYBRID SEED # 35A19, purchased from Pioneer Hi-Bred

International, Inc., Supply Management, P.O. Box 256, Johnston, Iowa 50131-0256.

A number of the nucleotide sequences disclosed in TABLE 1 herein as representative

of the SDFs of the invention can also be obtained by sequencing genomic DNA from

Arabidopsis thaliana, Wassilewskija ecotype or by sequencing cDNA obtained from mRNA

25   from such plants as described below. This is a true breeding strain. Seeds of the plant are

available from the Arabidopsis Biological Resource Center at the Ohio State University,

under the accession number CS2360. Seeds of this plant were deposited under the terms and

conditions of the Budapest Treaty at the American Type Culture Collection, Manassas, VA

on August 31, 1999, and were assigned ATCC No. PTA-595.

30   Other methods for cloning full-length cDNA are described, for example, by Seki et

al., *Plant Journal* 15:707-720 (1998) High-efficiency cloning of Arabidopsis full-length

cDNA by biotinylated Cap trapper"; Maruyama et al., *Gene* 138:171 (1994) Oligo-capping a

simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides";
and WO 96/34981.

Tissues were, or each organ was, individually pulverized and frozen in liquid
nitrogen. Next, the samples were homogenized in the presence of detergents and then
5   centrifuged. The debris and nuclei were removed from the sample and more detergents were
added to the sample. The sample was centrifuged and the debris was removed. Then the
sample was applied to a 2M sucrose cushion to isolate polysomes. The RNA was isolated by
treatment with detergents and proteinase K followed by ethanol precipitation and
centrifugation. The polysomal RNA from the different tissues was pooled according to the
10   following mass ratios: 15/15/1 for male inflorescences, female inflorescences and root,
respectively. The pooled material was then used for cDNA synthesis by the methods
described below.

Starting material for cDNA synthesis for the exemplary corn cDNA clones
with sequences presented in TABLE 1 was poly(A)-containing polysomal mRNAs from
15   inflorescences and root tissues of corn plants grown from HYBRID SEED # 35A19. Male
inflorescences and female (pre-and post-fertilization) inflorescences were isolated at various
stages of development. Selection for poly(A) containing polysomal RNA was done using
oligo d(T) cellulose columns, as described by Cox and Goldberg, Plant Molecular Biology:
A Practical Approach", pp. 1-35, Shaw ed., c. 1988 by IRL, Oxford. The quality and the
20   integrity of the polyA+ RNAs were evaluated.

Starting material for cDNA synthesis for the exemplary *Arabidopsis* cDNA
clones with sequences presented in TABLE 1 was polysomal RNA isolated from the top-
most inflorescence tissues of *Arabidopsis thaliana* Wassilewskija (Ws.) and from roots of
25   *Arabidopsis thaliana* Landsberg erecta (L. er.), also obtained from the Arabidopsis
Biological Resource Center. Nine parts inflorescence to every part root was used, as
measured by wet mass. Tissue was pulverized and exposed to liquid nitrogen. Next, the
sample was homogenized in the presence of detergents and then centrifuged. The debris and
nuclei were removed from the sample and more detergents were added to the sample. The
30   sample was centrifuged and the debris was removed and the sample was applied to a 2M
sucrose cushion to isolate polysomal RNA. Cox et al., Plant Molecular Biology: A Practical
Approach", pp. 1-35, Shaw ed., c. 1988 by IRL, Oxford. The polysomal RNA was used
for cDNA synthesis by the methods described below. Polysomal mRNA was then isolated as
described above for corn cDNA. The quality of the RNA was assessed electrophoretically.

Following preparation of the mRNAs from various tissues as described above, selection of mRNA with intact 5' ends and specific attachment of an oligonucleotide tag to the 5' end of such mRNA was performed using either a chemical or enzymatic approach. Both techniques take advantage of the presence of the cap" structure, which characterizes the 5' end of most

5 intact mRNAs and which comprises a guanosine generally methylated once, at the 7 position.

The chemical modification approach involves the optional elimination of the 2', 3'-cis diol of the 3' terminal ribose, the oxidation of the 2', 3'-cis diol of the ribose linked to the cap of the 5' ends of the mRNAs into a dialdehyde, and the coupling of the such obtained dialdehyde to a derivatized oligonucleotide tag. Further detail regarding the chemical approaches for

10 obtaining mRNAs having intact 5' ends are disclosed in International Application No. WO96/34981 published November 7, 1996.

The enzymatic approach for ligating the oligonucleotide tag to the intact 5' ends of mRNAs involves the removal of the phosphate groups present on the 5' ends of uncapped incomplete mRNAs, the subsequent decapping of mRNAs having intact 5' ends and the ligation

15 of the phosphate present at the 5' end of the decapped mRNA to an oligonucleotide tag. Further detail regarding the enzymatic approaches for obtaining mRNAs having intact 5' ends are disclosed in Dumas Milne Edwards J.B. (Doctoral Thesis of Paris VI University, Le clonage des ADNc complets: difficultés et perspectives nouvelles. Apports pour l'étude de la régulation de l'expression de la tryptophane hydroxylase de rat, 20 Dec. 1993), EP0 625572 and Kato *et al.*,

20 *Gene* 150:243-250 (1994).

In both the chemical and the enzymatic approach, the oligonucleotide tag has a restriction enzyme site (e.g. an EcoRI site) therein to facilitate later cloning procedures. Following attachment of the oligonucleotide tag to the mRNA, the integrity of the mRNA is examined by performing a Northern blot using a probe complementary to the oligonucleotide

25 tag.

For the mRNAs joined to oligonucleotide tags using either the chemical or the enzymatic method, first strand cDNA synthesis is performed using an oligo-dT primer with reverse transcriptase. This oligo-dT primer can contain an internal tag of at least 4 nucleotides, which can be different from one mRNA preparation to another. Methylated dCTP is used for cDNA

30 first strand synthesis to protect the internal EcoRI sites from digestion during subsequent steps. The first strand cDNA is precipitated using isopropanol after removal of RNA by alkaline hydrolysis to eliminate residual primers.

Second strand cDNA synthesis is conducted using a DNA polymerase, such as Klenow fragment and a primer corresponding to the 5' end of the ligated oligonucleotide. The primer is

typically 20-25 bases in length. Methylated dCTP is used for second strand synthesis in order to protect internal EcoRI sites in the cDNA from digestion during the cloning process.

Following second strand synthesis, the full-length cDNAs are cloned into a phagemid vector, such as pBlueScript™ (Stratagene). The ends of the full-length cDNAs are blunted with

5   T4 DNA polymerase (Biolabs) and the cDNA is digested with EcoRI. Since methylated dCTP is used during cDNA synthesis, the EcoRI site present in the tag is the only hemi-methylated site; hence the only site susceptible to EcoRI digestion. In some instances, to facilitate subcloning, an Hind III adapter is added to the 3' end of full-length cDNAs.

The full-length cDNAs are then size fractionated using either exclusion chromatography

10   (AcA, Biosepra) or electrophoretic separation which yields 3 to 6 different fractions. The full-length cDNAs are then directionally cloned either into pBlueScript™ using either the EcoRI and SmaI restriction sites or, when the Hind III adapter is present in the full-length cDNAs, the EcoRI and Hind III restriction sites. The ligation mixture is transformed, preferably by electroporation, into bacteria, which are then propagated under appropriate antibiotic selection.

15   Clones containing the oligonucleotide tag attached to full-length cDNAs are selected as follows.

The plasmid cDNA libraries made as described above are purified (e.g. by a column available from Qiagen). A positive selection of the tagged clones is performed as follows. Briefly, in this selection procedure, the plasmid DNA is converted to single stranded DNA using

20   phage F1 gene II endonuclease in combination with an exonuclease (Chang et al., Gene 127:95 (1993)) such as exonuclease III or T7 gene 6 exonuclease. The resulting single stranded DNA is then purified using paramagnetic beads as described by Fry et al., Biotechniques 13: 124 (1992). Here the single stranded DNA is hybridized with a biotinylated oligonucleotide having a sequence corresponding to the 3' end of the oligonucleotide tag. Preferably, the primer has a

25   length of 20-25 bases. Clones including a sequence complementary to the biotinylated oligonucleotide are selected by incubation with streptavidin coated magnetic beads followed by magnetic capture. After capture of the positive clones, the plasmid DNA is released from the magnetic beads and converted into double stranded DNA using a DNA polymerase such as ThermoSequenase™ (obtained from Amersham Pharmacia Biotech). Alternatively, protocols

30   such as the Gene Trapper™ kit (Gibco BRL) can be used. The double stranded DNA is then transformed, preferably by electroporation, into bacteria. The percentage of positive clones having the 5' tag oligonucleotide is typically estimated to be between 90 and 98% from dot blot analysis.

Following transformation, the libraries are ordered in microtiter plates and sequenced. The *Arabidopsis* library was deposited at the American Type Culture Collection on January 7, 2000 as *E-coli* liba 010600" under the accession number **PTA-1161.**

EXAMPLE 2: SOUTHERN HYBRIDIZATIONS

5   The SDFs of the invention can be used in Southern hybridizations as described above. The following describes extraction of DNA from nuclei of plant cells, digestion of the nuclear DNA and separation by length, transfer of the separated fragments to membranes, preparation of probes for hybridization, hybridization and detection of the hybridized probe.

The procedures described herein can be used to isolate related polynucleotides or for

10   diagnostic purposes. Moderate stringency hybridization conditions, as defined above, are described in the present example. These conditions result in detection of hybridization between sequences having at least 70% sequence identity. As described above, the hybridization and wash conditions can be changed to reflect the desired percenatge of sequence identity between probe and target sequences that can be detected.

15   In the following procedure, a probe for hybridization is produced from two PCR reactions using two primers from genomic sequence of *Arabidopsis thaliana*. As described above, the particular template for generating the probe can be any desired template.

The first PCR product is assessed to validate the size of the primer to assure it is of the expected size. Then the product of the first PCR is used as a template, with the same pair

20   of primers used in the first PCR, in a second PCR that produces a labeled product used as the probe.

Fragments detected by hybridization, or other bands of interest, can be isolated from gels used to separate genomic DNA fragments by known methods for further purification and/or characterization.

25   **Buffers for nuclear DNA extraction**

1.   10X HB

	1000 ml	
40 mM spermidine	10.2 g	Spermine (Sigma S-2876) and spermidine (Sigma S-2501)
10 mM spermine	3.5 g	Stabilize chromatin and the nuclear membrane

0.1 M EDTA (disodium)	37.2 g	EDTA inhibits nuclease
0.1 M Tris	12.1 g	Buffer
0.8 M KCl	59.6 g	Adjusts ionic strength for stability of nuclei

Adjust pH to 9.5 with 10 N NaOH. It appears that there is a nuclease present in leaves. Use of pH 9.5 appears to inactivate this nuclease.

2. 2 M sucrose (684 g per 1000 ml)

Heat about half the final volume of water to about 50°C. Add the sucrose slowly then bring the mixture to close to final volume; stir constantly until it has dissolved. Bring the solution to volume.

3. Sarkosyl solution (lyses nuclear membranes)

	1000 ml
N-lauroyl sarcosine (Sarkosyl)	20.0 g
0.1 M Tris	12.1 g
0.04 M EDTA (Disodium) 14.9 g	

Adjust the pH to 9.5 after all the components are dissolved and bring up to the proper volume.

4. 20% Triton X-100

80 ml Triton X-100

320 ml 1xHB (w/o β-ME and PMSF)

Prepare in advance; Triton takes some time to dissolve

A. Procedure

1. Prepare 1X H" buffer (keep ice-cold during use)

829

	<u>1000 ml</u>
10X HB	100 ml
2 M sucrose	250 ml a non-ionic osmoticum
Water	634 ml

5   **Added just before use:**

100 mM PMSF*	10 ml a protease inhibitor; protects nuclear membrane proteins
ß-mercaptoethanol	1 ml inactivates nuclease by reducing disulfide bonds

10   *100 mM PMSF

(phenyl methyl sulfonyl fluoride, Sigma P-7626)

(add 0.0875 g to 5 ml 100% ethanol)

2.   Homogenize the tissue in a blender (use 300-400 ml of 1xHB per blender). Be sure that you use 5-10 ml of HB buffer per gram of tissue.   Blenders generate heat so be

15   sure to keep the homogenate cold. It is necessary to put the blenders in ice periodically.

3.   Add the 20% Triton X-100 (25 ml per liter of homogenate) and gently stir on ice for 20 min. This lyses plastid, but not nuclear, membranes.

4.   Filter the tissue suspension through several nylon filters into an ice-cold beaker. The

20   first filtration is through a 250-micron membrane; the second is through an 85-micron membrane; the third is through a 50-micron membrane; and the fourth is through a 20-micron membrane. Use a large funnel to hold the filters. Filtration can be sped up by gently squeezing the liquid through the filters.

5.   Centrifuge the filtrate at 1200 x g for 20 min. at 4°C to pellet the nuclei.

6.   Discard the dark green supernatant. The pellet will have several layers to it. One is starch; it is white and gritty. The nuclei are gray and soft. In the early steps, there may be a dark green and somewhat viscous layer of chloroplasts.

Wash the pellets in about 25 ml cold H buffer (with Triton X-100) and resuspend by swirling gently and pipetting. After the pellets are resuspended.

5

Pellet the nuclei again at 1200 - 1300 x g. Discard the supernatant.

Repeat the wash 3-4 times until the supernatant has changed from a dark green to a pale green. This usually happens after 3 or 4 resuspensions. At this point, the pellet is typically grayish white and very slippery. The Triton X-100 in these repeated steps helps to destroy the chloroplasts and mitochondria that contaminate the prep.

10

Resuspend the nuclei for a final time in a total of 15 ml of H buffer and transfer the suspension to a sterile 125 ml Erlenmeyer flask.

7.   Add 15 ml, dropwise, cold 2% Sarkosyl, 0.1 M Tris, 0.04 M EDTA solution (pH 9.5) while swirling gently. This lyses the nuclei. The solution will become very viscous.

15   8.   Add 30 grams of CsCl and gently swirl at room temperature until the CsCl is in solution. The mixture will be gray, white and viscous.

9.   Centrifuge the solution at 11,400 x g at 4°C for at least 30 min. The longer this spin is, the firmer the protein pellicle.

20   10.   The result is typically a clear green supernatant over a white pellet, and (perhaps) under a protein pellicle. Carefully remove the solution under the protein pellicle and above the pellet. Determine the density of the solution by weighing 1 ml of solution and add CsCl if necessary to bring to 1.57 g/ml. The solution contains dissolved solids (sucrose etc) and the refractive index alone will not be an accurate guide to

25   CsCl concentration.

11.   Add 20 μl of 10 mg/ml EtBr per ml of solution.

12.   Centrifuge at 184,000 x g for 16 to 20 hours in a fixed-angle rotor.

13.   Remove the dark red supernatant that is at the top of the tube with a plastic transfer pipette and discard. Carefully remove the DNA band with another transfer pipette. The DNA band is usually visible in room light; otherwise, use a long wave UV light to locate the band.

14.   Extract the ethidium bromide with isopropanol saturated with water and salt. Once the solution is clear, extract at least two more times to ensure that all of the EtBr is gone. Be very gentle, as it is very easy to shear the DNA at this step. This extraction may take a while because the DNA solution tends to be very viscous. If the solution is too viscous, dilute it with TE.

15.   Dialyze the DNA for at least two days against several changes (at least three times) of TE (10 mM Tris, 1mM EDTA, pH 8) to remove the cesium chloride.

16.   Remove the dialyzed DNA from the tubing. If the dialyzed DNA solution contains a lot of debris, centrifuge the DNA solution at least at 2500 x g for 10 min. and carefully transfer the clear supernatant to a new tube. Read the A260 concentration of the DNA.

17.   Assess the quality of the DNA by agarose gel electrophoresis (1% agarose gel) of the DNA. Load 50 ng and 100 ng (based on the OD reading) and compare it with known and good quality DNA. Undigested lambda DNA and a lambda-HindIII-digested DNA are good molecular weight makers.

Protocol for Digestion of Genomic DNA

Protocol:

1.   The relative amounts of DNA for different crop plants that provide approximately a balanced number of genome equivalent is given in Table 3. Note that due to the size

of the wheat genome, wheat DNA will be underrepresented. Lambda DNA provides a useful control for complete digestion.

2.   Precipitate the DNA by adding 3 volumes of 100% ethanol. Incubate at -20°C for at least two hours. Yeast DNA can be purchased and made up at the necessary

5   concentration, therefore no precipitation is necessary for yeast DNA.

3.   Centrifuge the solution at 11,400 x g for 20 min. Decant the ethanol carefully (be careful not to disturb the pellet). Be sure that the residual ethanol is completely removed either by vacuum desiccation or by carefully wiping the sides of the tubes with a clean tissue.

10   4.   Resuspend the pellet in an appropriate volume of water. Be sure the pellet is fully resuspended before proceeding to the next step. This may take about 30 min.

5.   Add the appropriate volume of 10X reaction buffer provided by the manufacturer of the restriction enzyme to the resuspended DNA followed by the appropriate volume of enzymes. Be sure to mix it properly by slowly swirling the tubes.

15   6.   Set-up the lambda digestion-control for each DNA that you are digesting.

7.   Incubate both the experimental and lambda digests overnight at 37°C. Spin down condensation in a microfuge before proceeding.

8.   After digestion, add 2 µl of loading dye (typically 0.25% bromophenol blue, 0.25% xylene cyanol in 15% Ficoll or 30% glycerol) to the lambda-control digests and load

20   in 1% TPE-agarose gel (TPE is 90 mM Tris-phosphate, 2 mM EDTA, pH 8). If the lambda DNA in the lambda control digests are completely digested, proceed with the precipitation of the genomic DNA in the digests.

9.   Precipitate the digested DNA by adding 3 volumes of 100% ethanol and incubating in -20°C for at least 2 hours (preferably overnight).

EXCEPTION: *Arabidopsis* and yeast DNA are digested in an appropriate volume; they don't have to be precipitated.

10.   Resuspend the DNA in an appropriate volume of TE (e.g., 22 µl x 50 blots = 1100 µl) and an appropriate volume of 10X loading dye (e.g., 2.4 µl x 50 blots = 120 µl). Be careful in pipetting the loading dye - it is viscous. Be sure you are pipetting the correct volume.

Table 3

Some guide points in digesting genomic DNA.

Species	Genome Size	Size Relative to Arabidopsis	Genome Equivalent to 2 µg Arabidopsis DNA	Amount of DNA per blot
Arabidopsis	120 Mb	1X	1X	2 µg
Brassica	1,100 Mb	9.2X	0.54X	10 µg
Corn	2,800 Mb	23.3X	0.43X	20 µg
Cotton	2,300 Mb	19.2X	0.52X	20 µg
Oat	11,300 Mb	94X	0.11X	20 µg
Rice	400 Mb	3.3X	0.75X	5 µg
Soybean	1,100 Mb	9.2X	0.54X	10 µg
Sugarbeet	758 Mb	6.3X	0.8X	10 µg
Sweetclover	1,100 Mb	9.2X	0.54X	10 µg
Wheat	16,000 Mb	133X	0.08X	20 µg
Yeast	15 Mb	0.12X	1X	0.25 µg

Protocol for Southern Blot Analysis

The digested DNA samples are electrophoresed in 1% agarose gels in 1x TPE buffer. Low voltage; overnight separations are preferred. The gels are stained with EtBr and photographed.

1. For blotting the gels, first incubate the gel in 0.25 N HCl (with gentle shaking) for about 15 min.

2. Then briefly rinse with water. The DNA is denatured by 2 incubations. Incubate (with shaking) in 0.5 M NaOH in 1.5 M NaCl for 15 min.

5 3. The gel is then briefly rinsed in water and neutralized by incubating twice (with shaking) in 1.5 M Tris pH 7.5 in 1.5 M NaCl for 15 min.

4. A nylon membrane is prepared by soaking it in water for at least 5 min, then in 6X SSC for at least 15 min. before use. (20x SSC is 175.3 g NaCl, 88.2 g sodium citrate per liter, adjusted to pH 7.0.)

10 5. The nylon membrane is placed on top of the gel and all bubbles in between are removed. The DNA is blotted from the gel to the membrane using an absorbent medium, such as paper toweling and 6x SCC buffer. After the transfer, the membrane may be lightly brushed with a gloved hand to remove any agarose sticking to the surface.

15 6. The DNA is then fixed to the membrane by UV crosslinking and baking at 80°C. The membrane is stored at 4°C until use.

B. Protocol for PCR Amplification of Genomic Fragments in Arabidopsis

Amplification procedures:

1. Mix the following in a 0.20 ml PCR tube or 96-well PCR plate:

Volume	Stock	Final Amount or Conc.
0.5 μl	~ 10 ng/μl genomic DNA[1]	5 ng
2.5 μl	**10X PCR buffer**	20 mM Tris, 50 mM KCl

---

[1] Arabidopsis DNA is used in the present experiment, but the procedure is a general one.

0.75 µl	50 mM MgCl$_2$	1.5 mM
1 µl	10 pmol/µl Primer 1 (Forward)	10 pmol
1 µl	10 pmol/µl Primer 2 (Reverse)	10 pmol
0.5 µl	5 mM dNTPs	0.1 mM
0.1 µl	5 units/µl Platinum Taq™ (Life Technologies, Gaithersburg, MD) DNA Polymerase	1 units
(to 25 µl)	**Water**	

2.   The template DNA is amplified using a Perkin Elmer 9700 PCR machine:

1)   94°C for 10 min. followed by

2) 5 cycles:	3) 5 cycles:	4) 25 cycles:
94 °C - 30 sec 62 °C - 30 sec 72 °C - 3 min	94 °C - 30 sec 58 °C - 30 sec 72 °C - 3 min	94 °C - 30 sec 53 °C - 30 sec 72 °C - 3 min

5)   72°C for 7 min. Then the reactions are stopped by chilling to 4°C.

The procedure can be adapted to a multi-well format if necessary.

Quantification and Dilution of PCR Products:

1.   The product of the PCR is analyzed by electrophoresis in a 1% agarose gel. A linearized plasmid DNA can be used as a quantification standard (usually at 50, 100,

200, and 400 ng). These will be used as references to approximate the amount of PCR products. HindIII-digested Lambda DNA is useful as a molecular weight marker. The gel can be run fairly quickly; e.g., at 100 volts. The standard gel is examined to determine that the size of the PCR products is consistent with the expected size and if there are significant extra bands or smeary products in the PCR reactions.

2.   The amounts of PCR products can be estimated on the basis of the plasmid standard.

3.   For the small number of reactions that produce extraneous bands, a small amount of DNA from bands with the correct size can be isolated by dipping a sterile 10-μl tip into the band while viewing though a UV Transilluminator. The small amount of agarose gel (with the DNA fragment) is used in the labeling reaction.

C.   Protocol for PCR-DIG-Labeling of DNA

Solutions:

Reagents in PCR reactions (diluted PCR products, 10X PCR Buffer, 50 mM $MgCl_2$, 5 U/μl Platinum Taq Polymerase, and the primers)

10X dNTP + DIG-11-dUTP [1:5]:   (2 mM dATP, 2 mM dCTP, 2 mM dGTP, 1.65 mM dTTP, 0.35 mM DIG-11-dUTP)

10X dNTP + DIG-11-dUTP [1:10]:   (2 mM dATP, 2 mM dCTP, 2 mM dGTP, 1.81 mM dTTP, 0.19 mM DIG-11-dUTP)

10X dNTP + DIG-11-dUTP [1:15]:   (2 mM dATP, 2 mM dCTP, 2 mM dGTP, 1.875 mM dTTP, 0.125 mM DIG-11-dUTP)

TE buffer (10 mM Tris, 1 mM EDTA, pH 8)

Maleate buffer: In 700 ml of deionized distilled water, dissolve 11.61 g maleic acid and 8.77 g NaCl. Add NaOH to adjust the pH to 7.5. Bring the volume to 1 L. Stir for 15 min. and sterilize.

10% blocking solution: In 80 ml deionized distilled water, dissolve 1.16g maleic acid. Next, add NaOH to adjust the pH to 7.5. Add 10 g of the blocking reagent powder (Boehringer Mannheim, Indianapolis, IN, Cat. no. 1096176). Heat to 60°C while stirring to dissolve the powder. Adjust the volume to 100 ml with water. Stir and sterilize.

5

1% blocking solution: Dilute the 10% stock to 1% using the maleate buffer.

Buffer 3 (100 mM Tris, 100 mM NaCl, 50 mM $MgCl_2$, pH9.5). Prepared from autoclaved solutions of 1M Tris pH 9.5, 5 M NaCl, and 1 M $MgCl_2$ in autoclaved distilled water.

Procedure:

1. PCR reactions are performed in 25 μl volumes containing:

PCR buffer	1X
MgCl$_2$	1.5 mM
10X dNTP + DIG-11-dUTP	1X (please see the note below)
Platinum Taq™ Polymerase	1 unit
10 pg probe DNA	
10 pmol primer 1	

Note:	Use for:
10X dNTP + DIG-11-dUTP (1:5)	< 1 kb
10X dNTP + DIG-11-dUTP (1:10)	1 kb to 1.8 kb
10X dNTP + DIG-11-dUTP (1:15)	> 1.8 kb

2. The PCR reaction uses the following amplification cycles:

1) 94°C for 10 min.

2) 5 cycles:	3) 5 cycles:	4) 25 cycles:
95°C - 30 sec 61°C - 1 min 73°C - 5 min	95°C - 30 sec 59°C - 1 min 75°C - 5 min	95°C - 30 sec 51°C - 1 min 73°C - 5 min

5) 72°C for 8 min. The reactions are terminated by chilling to 4°C (hold).

3. The products are analyzed by electrophoresis- in a 1% agarose gel, comparing to an aliquot of the unlabelled probe starting material.

4. The amount of DIG-labeled probe is determined as follows:

Make serial dilutions of the diluted control DNA in dilution buffer (TE: 10 mM Tris and 1 mM EDTA, pH 8) as shown in the following table:

DIG-labeled control DNA starting conc.	Stepwise Dilution	Final Conc. (Dilution Name)
5 ng/μl	1 μl in 49 μl TE	100 pg/μl (A)
100 pg/μl (A)	25 μl in 25 μl TE	50 pg/μl (B)
50 pg/μl (B)	25 μl in 25 μl TE	25 pg/μl (C)
25 pg/μl (C)	20 μl in 30 μl TE	10 pg/μl (D)

a.   Serial deletions of a DIG-labeled standard DNA ranging from 100 pg to 10 pg are spotted onto a positively charged nylon membrane, marking the membrane lightly with a pencil to identify each dilution.

b.   Serial dilutions (e.g., 1:50, 1:2500, 1:10,000) of the newly labeled DNA probe are spotted.

c.   The membrane is fixed by UV crosslinking.

d.   The membrane is wetted with a small amount of maleate buffer and then incubated in 1% blocking solution for 15 min at room temp.

e.   The labeled DNA is then detected using alkaline phosphatase conjugated anti-DIG antibody (Boehringer Mannheim, Indianapolis, IN, cat. no. 1093274) and an NBT substrate according to the manufacture's instruction.

f.   Spot intensities of the control and experimental dilutions are then compared to estimate the concentration of the PCR-DIG-labeled probe.

D. Prehybridization and Hybridization of Southern Blots

Solutions:

100% Formamide   purchased from Gibco

20X SSC   (1X = 0.15 M NaCl, 0.015 M Na₃citrate)

per L: 175 g NaCl

87.5 g Na₃citrate·2H₂0

20% Sarkosyl (N-lauroyl-sarcosine)

20% SDS (sodium dodecyl sulphate)

10% Blocking Reagent: In 80 ml deionized distilled water, dissolve 1.16 g maleic acid. Next, add NaOH to adjust the pH to 7.5. Add 10 g of the blocking reagent powder. Heat to 60°C while stirring to dissolve the powder. Adjust the volume to 100 ml with water. Stir and sterilize.

Prehybridization Mix:

Final Concentration	Components	Volume (per 100 ml)	Stock
50%	Formamide	50 ml	100%
5X	SSC	25 ml	20X
0.1%	Sarkosyl	0.5 ml	20%
0.02%	SDS	0.1 ml	20%
2%	Blocking Reagent	20 ml	10%
	Water	4.4 ml	

General Procedures:

1. Place the blot in a heat-sealable plastic bag and add an appropriate volume of prehybridization solution (30 ml/100cm^2) at room temperature. Seal the bag with a heat sealer, avoiding bubbles as much as possible. Lay down the bags in a large plastic tray (one tray can accommodate at least 4–5 bags). Ensure that the bags are

lying flat in the tray so that the prehybridization solution is evenly distributed throughout the bag. Incubate the blot for at least 2 hours with gentle agitation using a waver shaker.

2. Denature DIG-labeled DNA probe by incubating for 10 min. at 98°C using the PCR machine and immediately cool it to 4°C.

3. Add probe to prehybridization solution (25 ng/ml; 30 ml = 750 ng total probe) and mix well but avoid foaming. Bubbles may lead to background.

4. Pour off the prehybridization solution from the hybridization bags and add new prehybridization and probe solution mixture to the bags containing the membrane.

5. Incubate with gentle agitation for at least 16 hours.

6. Proceed to medium stringency post-hybridization wash:

Three times for 20 min. each with gentle agitation using 1X SSC, 1% SDS at 60°C.

All wash solutions must be prewarmed to 60°C. Use about 100 ml of wash solution per membrane.

To avoid background keep the membranes fully submerged to avoid drying in spots; agitate sufficiently to avoid having membranes stick to one another.

7. After the wash, proceed to immunological detection and CSPD development.

E.   **Procedure for Immunological Detection with CSPD**
Solutions:

Buffer 1:	Maleic acid buffer (0.1 M maleic acid, 0.15 M NaCl; adjusted to pH 7.5 with NaoH)
Washing buffer:	Maleic acid buffer with 0.3% (v/v) Tween 20.

Blocking stock solution

10% blocking reagent in buffer 1. Dissolve (10X concentration): blocking reagent powder (Boehringer Mannheim, Indianapolis, IN, cat. no. 1096176) by constantly stirring on a 65°C heating block or heat in a microwave, autoclave and store at 4°C.

Buffer 2

(1X blocking solution):   Dilute the stock solution 1:10 in Buffer 1.

Detection buffer: 0.1 M Tris, 0.1 M NaCl, pH 9.5

Procedure:

1.   After the post-hybridization wash the blots are briefly rinsed (1-5 min.) in the maleate washing buffer with gentle shaking.

2.   Then the membranes are incubated for 30 min. in Buffer 2 with gentle shaking.

3.   Anti-DIG-AP conjugate (Boehringer Mannheim, Indianapolis, IN, cat. no. 1093274) at 75 mU/ml (1:10,000) in Buffer 2 is used for detection. 75 ml of solution can be used for 3 blots.

4.   The membrane is incubated for 30 min. in the antibody solution with gentle shaking.

5.   The membrane are washed twice in washing buffer with gentle shaking. About 250 mls is used per wash for 3 blots.

6.   The blots are equilibrated for 2–5 min in 60 ml detection buffer.

7.   Dilute CSPD (1:200) in detection buffer. (This can be prepared ahead of time and stored in the dark at 4°C).

The following steps must be done individually. Bags (one for detection and one for exposure) are generally cut and ready before doing the following steps.

8.   The blot is carefully removed from the detection buffer and excess liquid removed
without drying the membrane. The blot is immediately placed in a bag and 1.5 ml of
CSPD solution is added. The CSPD solution can be spread over the membrane.
Bubbles present at the edge and on the surface of the blot are typically removed by
5   gentle rubbing. The membrane is incubated for 5 min. in CSPD solution.

9.   Excess liquid is removed and the membrane is blotted briefly (DNA side up) on
Whatman 3MM paper. Do not let the membrane dry completely.

10.   Seal the damp membrane in a hybridization bag and incubate for 10 min at 37°C to
enhance the luminescent reaction.

10   11.   Expose for 2 hours at room temperature to X-ray film. Multiple exposures can be
taken. Luminescence continues for at least 24 hours and signal intensity increases
during the first hours.

Example 3: Transformation of Carrot Cells

        Transformation of plant cells can be accomplished by a number of methods, as
15   described above. Similarly, a number of plant genera can be regenerated from tissue culture
following transformation. Transformation and regeneration of carrot cells as described herein
is illustrative.

        Single cell suspension cultures of carrot (*Daucus carota*) cells are established from
hypocotyls of cultivar Early Nantes in B_5 growth medium (O.L. Gamborg et al., *Plant*
20   *Physiol.* 45:372 (1970)) plus 2,4-D and 15 mM $CaCl_2$ (B_5 -44 medium) by methods known in
the art. The suspension cultures are subcultured by adding 10 ml of the suspension culture to
40 ml of B_5-44 medium in 250 ml flasks every 7 days and are maintained in a shaker at 150
rpm at 27 $^{\circ}$C in the dark.

        The suspension culture cells are transformed with exogenous DNA as described by Z.
25   Chen et al. *Plant Mol. Bio.* 36:163 (1998). Briefly, 4-days post-subculture cells are incubated
with cell wall digestion solution containing 0.4 M sorbitol, 2% driselase, 5mM MES (2-[N-
Morpholino] ethanesulfonic acid) pH 5.0 for 5 hours. The digested cells are pelleted gently
at 60 xg for 5 min. and washed twice in W5 solution containing 154 mM NaCl, 5 mM KCl,
125 mM $CaCl_2$ and 5mM glucose, pH 6.0. The protoplasts are suspended in MC solution

containing 5 mM MES, 20 mM CaCl$_2$, 0.5 M mannitol, pH 5.7 and the protoplast density is adjusted to about 4 x 10^6 protoplasts per ml.

15-60 µg of plasmid DNA is mixed with 0.9 ml of protoplasts. The resulting suspension is mixed with 40% polyethylene glycol (MW 8000, PEG 8000), by gentle

5 inversion a few times at room temperature for 5 to 25 min. Protoplast culture medium known in the art is added into the PEG-DNA-protoplast mixture. Protoplasts are incubated in the culture medium for 24 hour to 5 days and cell extracts can be used for assay of transient expression of the introduced gene. Alternatively, transformed cells can be used to produce transgenic callus, which in turn can be used to produce transgenic plants, by methods known

10 in the art. See, for example, Nomura and Komamine, *Plt. Phys.* _79_:988-991 (1985), *Identification and Isolation of Single Cells that Produce Somatic Embryos in Carrot Suspension Cultures.*

An additional deposit, PTA-1411, of an *E. coli* Library, *E. coli*LibA021800, was made at the American Type Culture Collection in Manassas, Virginia, USA on February 22,

15 2000 to meet the requirements of Budapest Treaty for the international recognition of the deposit of microorganisms. This deposit was assigned ATCC accession no. PTA-1411.

The invention being thus described, it will be apparent to one of ordinary skill in the art that various modifications of the materials and methods for practicing the invention can be made. Such modifications are to be considered within the scope of the invention as defined

20 by the following claims.

Each of the references from the patent and periodical literature cited herein is hereby expressly incorporated in its entirety by such citation.

Table 1

>5541651   /

	len =	37560	nex =	61		
5						
	Term	351	143		–	0
	Intr	795	738		–	0
	Intr	971	922		–	0
	Intr	1292	1086		–	0
10	Intr	2075	1655		–	0
	Intr	2936	2841		–	0
	Init	3468	3148		–	0
	Term	8256	7580		–	1
15	Intr	8495	8342		–	1
	Intr	10104	8596		–	1
	Intr	10243	10200		–	1
	Init	10453	10327		–	1
20	Sngl	15598	15326		–	2
	Sngl	17736	16462		–	3
	Term	24957	24846		–	4
25	Intr	25196	25114		–	4
	Intr	25380	25300		–	4
	Intr	25525	25469		–	4
	Intr	25747	25691		–	4
	Intr	25939	25843		–	4
30	Intr	26088	26018		–	4
	Intr	26302	26174		–	4
	Intr	26598	26398		–	4
	Intr	26854	26702		–	4
	Intr	27206	27036		–	4
35	Intr	27351	27289		–	4
	Intr	27701	27590		–	4
	Intr	27930	27791		–	4
	Intr	28293	28081		–	4
	Intr	28517	28419		–	4
40	Intr	28759	28640		–	4
	Intr	29124	28919		–	4
	Intr	29419	29242		–	4
	Intr	29567	29507		–	4
	Intr	29754	29645		–	4
45	Intr	29980	29849		–	4
	Intr	30235	30068		–	4
	Intr	30512	30353		–	4
	Intr	30670	30633		–	4
	Intr	31006	30905		–	4
50	Intr	31612	31501		–	4
	Intr	31870	31736		–	4
	Intr	32110	31974		–	4
	Intr	32355	32206		–	4
	Intr	32640	32481		–	4
55	Intr	32775	32717		–	4
	Intr	33232	33076		–	4
	Intr	33480	33335		–	4
	Intr	33698	33555		–	4
	Intr	33927	33799		–	4
60	Intr	34294	34256		–	4

846

	Intr	34557	34501	–	4
	Intr	35392	35340	–	4
	Intr	35592	35466	–	4
	Intr	35779	35679	–	4
5	Intr	36117	35991	–	4
	Intr	36609	36499	–	4
	Intr	36851	36701	–	4
	Intr	37011	36959	–	4
	Intr	37176	37102	–	4
10	Init	37390	37313	–	4

>5541653   /

len = 79320   nex =   127

15	Sngl	203	215	+	0
	Init	2597	3003	+	1
	Term	3394	3721	+	1
20	Sngl	4128	4322	+	2
	Init	4924	5133	+	3
	Intr	5240	5302	+	3
25	Intr	5391	5672	+	3
	Intr	5800	5865	+	3
	Intr	5974	6102	+	3
	Intr	6231	6524	+	3
	Intr	6765	6827	+	3
30	Intr	6972	7049	+	3
	Intr	7141	7209	+	3
	Intr	7309	7392	+	3
	Intr	7531	7587	+	3
	Intr	7671	7747	+	3
35	Intr	7949	8030	+	3
	Intr	8121	8237	+	3
	Intr	8550	8683	+	3
	Intr	8852	8921	+	3
	Term	9024	9092	+	3
40	Sngl	11309	9474	–	4
	Init	11464	11808	+	5
	Intr	12005	12139	+	5
45	Intr	12237	12428	+	5
	Intr	12541	12602	+	5
	Intr	12725	12885	+	5
	Intr	12967	13115	+	5
	Intr	13194	13334	+	5
50	Intr	13416	13568	+	5
	Intr	14271	16052	+	5
	Intr	16442	16503	+	5
	Intr	16783	16967	+	5
	Intr	17266	17648	+	5
55	Intr	17732	18016	+	5
	Intr	18112	18347	+	5
	Term	18432	18609	+	5
	Init	21616	21783	+	6
60	Intr	21972	22366	+	6
	Intr	22453	22537	+	6

847

	Intr	22634	22856	+	6
	Intr	22944	23047	+	6
	Intr	23131	23221	+	6
	Intr	23328	23482	+	6
5	Intr	23656	24657	+	6
	Intr	24741	25127	+	6
	Term	25212	25508	+	6
	Term	26400	26248	−	7
10	Intr	26591	26485	−	7
	Intr	26878	26674	−	7
	Intr	27285	26952	−	7
	Intr	27556	27372	−	7
	Intr	27832	27637	−	7
15	Init	28924	27912	−	7
	Term	32372	32284	−	8
	Intr	32622	32480	−	8
	Intr	32841	32719	−	8
20	Intr	33040	32941	−	8
	Intr	33217	33135	−	8
	Intr	33471	33319	−	8
	Init	33983	33634	−	8
25	Sngl	35765	35103	−	9
	Term	36432	35963	−	10
	Intr	36698	36497	−	10
	Intr	37001	36902	−	10
30	Intr	37282	37196	−	10
	Intr	37647	37406	−	10
	Intr	37914	37832	−	10
	Intr	38349	38232	−	10
	Intr	38964	38576	−	10
35	Init	40843	39145	−	10
	Init	43542	43562	+	11
	Intr	43667	43867	+	11
	Intr	44142	44291	+	11
40	Intr	44386	44571	+	11
	Intr	44663	44743	+	11
	Intr	44910	45110	+	11
	Intr	45226	45366	+	11
	Term	45456	45653	+	11
45					
	Init	46353	46741	+	12
	Intr	46913	47010	+	12
	Intr	47095	47160	+	12
	Term	47519	48153	+	12
50					
	Sngl	48876	50090	+	13
	Init	52171	52413	+	14
	Intr	52540	52713	+	14
55	Intr	52865	53033	+	14
	Intr	53180	53241	+	14
	Intr	53331	53452	+	14
	Intr	53623	53740	+	14
	Term	53836	53973	+	14
60					
	Sngl	58340	59602	+	15

	Term	60245	60133	−	16
	Intr	60439	60334	−	16
	Intr	61880	61519	−	16
5	Init	62208	61986	−	16
	Term	63156	62911	−	17
	Intr	63595	63254	−	17
	Intr	63776	63660	−	17
10	Intr	64014	63850	−	17
	Init	64788	64612	−	17
	Init	65271	65491	+	18
	Term	65625	66057	+	18
15					
	Term	66668	66504	−	19
	Intr	66933	66778	−	19
	Intr	67142	67044	−	19
	Intr	67570	67248	−	19
20	Init	68343	67659	−	19
	Sngl	69081	68911	−	20
	Init	70351	70437	+	21
25	Intr	70537	70606	+	21
	Intr	70706	70784	+	21
	Intr	71159	71240	+	21
	Intr	71441	71549	+	21
	Intr	71688	71776	+	21
30	Intr	71853	71917	+	21
	Intr	72034	72112	+	21
	Term	72192	72359	+	21
	Init	72879	73109	+	22
35	Intr	73419	73646	+	22
	Intr	73748	73821	+	22
	Intr	73963	74122	+	22
	Intr	74217	74336	+	22
	Intr	74453	74561	+	22
40	Intr	74724	74879	+	22
	Term	74965	75032	+	22
	.Sngl	78028	79311	+	23
45	>5541654	/			
	len =	58920	nex =	63	
	Init	58	244	+	0
50	Term	3227	3330	+	0
	Init	5408	5527	+	1
	Intr	5889	5940	+	1
	Intr	6067	6147	+	1
55	Intr	6238	6356	+	1
	Intr	6668	6887	+	1
	Intr	6973	7091	+	1
	Term	7285	8040	+	1
60	Sngl	8735	9310	+	2

849

	Term	9809	9640	−	3
	Init	10251	10188	−	3
	Sngl	10639	10821	+	4
5					
	Term	12075	11900	−	5
	Init	12465	12402	−	5
	Sngl	13513	14094	+	6
10					
	Init	15064	15604	+	7
	Intr	16038	16108	+	7
	Intr	16260	18224	+	7
	Term	19124	19150	+	7
15					
	Init	21109	21501	+	8
	Intr	21566	22902	+	8
	Intr	22974	23084	+	8
	Intr	23177	23453	+	8
20	Intr	23543	23631	+	8
	Intr	23713	23821	+	8
	Intr	23906	24130	+	8
	Intr	24199	24324	+	8
	Intr	24399	24497	+	8
25	Intr	24573	24764	+	8
	Term	24836	25447	+	8
	Init	28588	29266	+	9
	Intr	30133	30823	+	9
30	Intr	30902	31263	+	9
	Intr	31288	31551	+	9
	Intr	31584	33236	+	9
	Intr	33292	33923	+	9
	Intr	34136	34828	+	9
35	Term	35080	35628	+	9
	Term	37540	37491	−	10
	Intr	37732	37643	−	10
	Init	38328	38238	−	10
40					
	Sngl	39544	39741	+	11
	Term	41267	40966	−	12
	Intr	41929	41738	−	12
45	Intr	42087	42001	−	12
	Intr	42553	42295	−	12
	Init	43798	43166	−	12
	Sngl	45500	45108	−	13
50					
	Sngl	46583	46020	−	14
	Term	48039	47391	−	15
	Init	48546	48257	−	15
55					
	Term	52521	52104	−	16
	Intr	53417	53288	−	16
	Init	53665	53527	−	16
60	Term	56895	56818	−	17
	Intr	57100	56966	−	17

850

	Intr	57356	57204	–	17
	Intr	57673	57475	–	17
	Intr	57938	57901	–	17
	Intr	58231	58060	–	17
5	Intr	58512	58390	–	17
	Init	58838	58602	–	17

>5541655   /

10	len =	36840	nex =	69		
	Init	1223	1228		+	0
	Intr	1302	1380		+	0
	Term	1441	1901		+	0
15						
	Init	2569	2832		+	1
	Intr	2932	3498		+	1
	Intr	3578	3757		+	1
	Intr	3917	4020		+	1
20	Intr	4094	4205		+	1
	Intr	4285	4383		+	1
	Intr	4471	4590		+	1
	Term	4684	4800		+	1
25	Term	7356	7204		–	2
	Intr	7705	7439		–	2
	Intr	7993	7787		–	2
	Intr	8251	8093		–	2
	Intr	8487	8350		–	2
30	Intr	8679	8574		–	2
	Intr	8932	8765		–	2
	Intr	9244	9054		–	2
	Intr	9660	9436		–	2
	Intr	9933	9816		–	2
35	Intr	10187	10024		–	2
	Intr	10381	10271		–	2
	Intr	10589	10484		–	2
	Intr	10872	10767		–	2
	Intr	11196	11026		–	2
40	Intr	11402	11281		–	2
	Intr	11635	11501		–	2
	Intr	11799	11719		–	2
	Intr	12092	11881		–	2
	Intr	12645	12156		–	2
45	Intr	13346	13240		–	2
	Init	16970	14193		–	2
	Init	18232	18596		+	3
	Intr	18799	19018		+	3
50	Intr	19099	19420		+	3
	Intr	19512	19906		+	3
	Intr	20122	20402		+	3
	Intr	20471	20502		+	3
	Intr	20852	20983		+	3
55	Intr	21370	21420		+	3
	Term	21557	21609		+	3
	Init	23028	23370		+	4
	Intr	23451	24565		+	4
60	Term	24656	24730		+	4

851

	Sngl	25809	26408	+	5
	Init	27959	29286	+	6
	Intr	29376	29542	+	6
5	Intr	29628	29866	+	6
	Term	29960	30574	+	6
	Term	31091	30909	−	7
	Intr	31376	31248	−	7
10	Intr	31936	31464	−	7
	Intr	32209	32032	−	7
	Intr	32622	32402	−	7
	Init	32755	32662	−	7
15	Term	33582	33472	−	8
	Intr	33809	33802	−	8
	Intr	34127	34028	−	8
	Intr	34256	34225	−	8
	Intr	34588	34371	−	8
20	Intr	34782	34667	−	8
	Intr	35122	34869	−	8
	Intr	35540	35465	−	8
	Intr	35690	35629	−	8
	Intr	35859	35799	−	8
25	Intr	35955	35921	−	8
	Init	36126	36033	−	8
	Sngl	36839	36687	−	9
30	>5541656	/			

len =   71580   nex =   90

	Sngl	137	140	+	0
35	Init	176	370	+	1
	Intr	467	736	+	1
	Term	992	1930	+	1
40	Init	4090	4284	+	2
	Intr	4333	4377	+	2
	Intr	4437	4706	+	2
	Term	4957	5904	+	2
45	Init	9007	9063	+	3
	Term	9243	9296	+	3
	Init	9716	9916	+	4
	Intr	10072	10512	+	4
50	Intr	10615	10884	+	4
	Term	10983	11930	+	4
	Init	13084	13278	+	5
	Intr	13375	13818	+	5
55	Intr	13908	14279	+	5
	Intr	14361	14583	+	5
	Intr	14996	15084	+	5
	Intr	15443	15505	+	5
	Intr	15833	15847	+	5
60	Intr	17101	17331	+	5
	Intr	17429	17872	+	5

852

	Term	17994	18245	+	5
	Init	19401	19601	+	6
	Intr	19703	20140	+	6
5	Term	20279	20530	+	6
	Sngl	21514	21699	+	7
	Init	23658	23852	+	8
10	Intr	24183	24638	+	8
	Term	24861	25112	+	8
	Init	26164	26358	+	9
	Intr	26543	26998	+	9
15	Term	27112	27363	+	9
	Init	30600	30746	+	10
	Intr	30850	31656	+	10
	Intr	31754	32206	+	10
20	Intr	32289	32750	+	10
	Term	32845	33093	+	10
	Term	33915	33667	−	11
	Intr	34455	34006	−	11
25	Intr	35002	34541	−	11
	Init	35280	35086	−	11
	Sngl	37170	37322	+	12
30	Init	37477	37515	+	13
	Intr	37805	38038	+	13
	Intr	38121	38417	+	13
	Intr	38504	38671	+	13
	Intr	38783	38950	+	13
35	Intr	39040	39117	+	13
	Intr	39191	39257	+	13
	Intr	39379	39476	+	13
	Intr	39566	39635	+	13
	Intr	39725	39795	+	13
40	Intr	39881	39930	+	13
	Intr	40024	40136	+	13
	Term	40241	40287	+	13
	Term	42156	41533	−	14
45	Intr	42426	42244	−	14
	Init	43060	42698	−	14
	Term	50406	50372	−	15
	Intr	50550	50483	−	15
50	Intr	50902	50770	−	15
	Intr	51275	50995	−	15
	Init	51779	51487	−	15
	Term	55279	54721	−	16
55	Init	55765	55395	−	16
	Term	56422	56418	−	17
	Intr	56753	56657	−	17
	Intr	57200	57037	−	17
60	Init	58404	57291	−	17

853

	Sngl	62987	62157	−	18
	Term	64148	63861	−	19
	Intr	64373	64323	−	19
5	Intr	64724	64596	−	19
	Intr	64936	64809	−	19
	Intr	65273	65014	−	19
	Intr	65562	65388	−	19
	Intr	65948	65706	−	19
10	Intr	66261	66020	−	19
	Intr	67497	67403	−	19
	Intr	67685	67575	−	19
	Intr	67832	67782	−	19
	Intr	68202	68056	−	19
15	Intr	68436	68311	−	19
	Intr	68778	68538	−	19
	Intr	69088	68893	−	19
	Intr	69642	69388	−	19
	Init	69988	69724	−	19
20					
	Sngl	70997	71323	+	20

>5541657　　　/

25	len =	20760	nex =	51		
	Init	889	991	+	0	
	Intr	1069	1164	+	0	
	Intr	1247	1302	+	0	
30	Intr	1398	1521	+	0	
	Intr	1611	1673	+	0	
	Intr	1774	1833	+	0	
	Intr	1933	2094	+	0	
	Intr	2173	2246	+	0	
35	Intr	2324	2417	+	0	
	Term	2500	2640	+	0	
	Init	5379	5521	+	1	
	Intr	5591	5690	+	1	
40	Intr	5772	5867	+	1	
	Intr	5948	6003	+	1	
	Intr	6096	6219	+	1	
	Intr	6272	6334	+	1	
	Intr	6376	6434	+	1	
45	Intr	6563	6749	+	1	
	Intr	6818	6891	+	1	
	Intr	6965	7058	+	1	
	Intr	7135	7271	+	1	
	Term	7490	7496	+	1	
50						
	Term	8030	7941	−	2	
	Intr	8384	8298	−	2	
	Intr	8568	8460	−	2	
	Intr	9171	9100	−	2	
55	Init	9510	9239	−	2	
	Init	13166	13342	+	3	
	Intr	13759	13818	+	3	
	Intr	14103	14426	+	3	
60	Intr	14519	14620	+	3	
	Intr	14691	14750	+	3	

854

	Intr	14842	15071	+	3
	Intr	15151	15274	+	3
	Intr	15365	15545	+	3
	Intr	15752	15773	+	3
5	Intr	15865	15900	+	3
	Intr	15974	16076	+	3
	Intr	16177	16272	+	3
	Intr	16373	16419	+	3
	Intr	16490	16613	+	3
10	Intr	16708	16770	+	3
	Intr	16865	17167	+	3
	Intr	17248	17321	+	3
	Intr	17407	17500	+	3
	Intr	17593	17729	+	3
15	Intr	17803	17883	+	3
	Intr	17968	18073	+	3
	Term	18174	18254	+	3
	Term	20027	19932	−	4
20	Init	20581	20524	−	4

>5541658   /

len =   63000   nex =   120

25					
	Init	1	131	+	0
	Term	167	229	+	0
	Term	1101	877	−	1
30	Intr	1361	1182	−	1
	Intr	1619	1503	−	1
	Intr	1833	1768	−	1
	Intr	2009	1918	−	1
	Intr	2250	2187	−	1
35	Intr	2430	2352	−	1
	Intr	2648	2509	−	1
	Intr	2804	2723	−	1
	Intr	3030	2927	−	1
	Intr	3270	3193	−	1
40	Intr	3516	3423	−	1
	Intr	3700	3603	−	1
	Init	4469	3912	−	1
	Term	5494	5385	−	2
45	Intr	5886	5583	−	2
	Intr	6288	5983	−	2
	Intr	6904	6395	−	2
	Intr	7184	7020	−	2
	Intr	7979	7272	−	2
50	Intr	8105	8019	−	2
	Intr	10111	8202	−	2
	Init	10616	10586	−	2
	Term	12494	12379	−	3
55	Intr	12819	12580	−	3
	Intr	12967	12904	−	3
	Intr	13621	13050	−	3
	Intr	13960	13666	−	3
	Intr	14239	14075	−	3
60	Intr	14962	14348	−	3
	Intr	15127	15041	−	3

855

	Init	17081	15240	-	3
	Term	18355	18225	-	4
	Intr	18679	18440	-	4
5	Intr	18828	18765	-	4
	Intr	19820	18912	-	4
	Intr	20166	20002	-	4
	Intr	20879	20268	-	4
	Intr	21046	20960	-	4
10	Intr	22578	21173	-	4
	Intr	22969	22699	-	4
	Intr	23073	23009	-	4
	Intr	24013	23926	-	4
	Init	25480	24314	-	4
15					
	Term	29987	29676	-	5
	Intr	30515	30477	-	5
	Init	30892	30668	-	5
20	Init	32168	32195	+	6
	Term	32330	32979	+	6
	Term	36474	36020	-	7
	Intr	36633	36558	-	7
25	Intr	36805	36723	-	7
	Init	38996	37385	-	7
	Sngl	40544	39360	-	8
30	Term	41055	40926	-	9
	Intr	41208	41145	-	9
	Intr	41425	41305	-	9
	Intr	41700	41572	-	9
	Intr	41804	41773	-	9
35	Intr	42090	42043	-	9
	Intr	42273	42237	-	9
	Intr	42418	42349	-	9
	Intr	42523	42507	-	9
	Intr	42699	42626	-	9
40	Intr	42947	42846	-	9
	Intr	43102	43029	-	9
	Intr	43374	43275	-	9
	Init	43558	43504	-	9
45	Term	44946	44626	-	10
	Intr	45823	45707	-	10
	Intr	45992	45896	-	10
	Intr	46187	46102	-	10
	Intr	46366	46295	-	10
50	Intr	46567	46449	-	10
	Intr	46806	46699	-	10
	Intr	47018	46958	-	10
	Intr	47181	47122	-	10
	Intr	47370	47310	-	10
55	Intr	47548	47457	-	10
	Intr	47866	47780	-	10
	Intr	48045	47962	-	10
	Intr	48314	48249	-	10
	Intr	48539	48426	-	10
60	Intr	48821	48667	-	10
	Intr	49102	49040	-	10

856

	Init	49332	49218	−	10
	Init	50117	50206	+	11
	Term	50310	51170	+	11
5					
	Init	51605	51735	+	12
	Intr	52114	52220	+	12
	Intr	52289	52330	+	12
	Intr	52425	52512	+	12
10	Intr	52786	52885	+	12
	Intr	53026	53126	+	12
	Term	53219	53342	+	12
	Init	54287	54622	+	13
15	Intr	54674	55422	+	13
	Intr	55674	55739	+	13
	Intr	55827	55971	+	13
	Intr	56061	56279	+	13
	Intr	56374	56572	+	13
20	Intr	56661	56856	+	13
	Term	56946	57063	+	13
	Term	57582	57334	−	14.
	Intr	57798	57691	−	14
25	Intr	58158	57874	−	14
	Intr	58526	58242	−	14
	Intr	58879	58598	−	14
	Intr	59149	58966	−	14
	Intr	59437	59282	−	14
30	Intr	59792	59503	−	14
	Init	60069	59872	−	14
	Term	60874	60653	−	15
	Intr	61357	60950	−	15
35	Intr	61667	61505	−	15
	Intr	62077	61798	−	15
	Intr	62421	62168	−	15
	Init	62629	62503	−	15
40	>5541660	/			
	len =	71220	nex =	95	
	Sngl	846	388	−	0
45					
	Init	1261	2018	+	1
	Term	2069	2339	+	1
	Term	2992	2955	−	2
50	Intr	3174	3078	−	2
	Intr	3418	3266	−	2
	Intr	3573	3502	−	2
	Intr	3791	3657	−	2
	Intr	4092	3892	−	2
55	Init	4671	4258	−	2
	Term	5794	5711	−	3
	Intr	5928	5878	−	3
	Intr	6091	6013	−	3
60	Intr	6482	6379	−	3
	Intr	6644	6563	−	3

857

	Intr	6846	6726	–	3
	Intr	7097	6938	–	3
	Intr	7231	7170	–	3
	Init	7764	7329	–	3
5					
	Init	8083	8634	+	4
	Intr	8722	8796	+	4
	Intr	9273	9359	+	4
	Intr	9394	9450	+	4
10	Intr	9648	9707	+	4
	Intr	9796	9876	+	4
	Intr	9968	10003	+	4
	Term	10090	10170	+	4
15	Init	11387	11607	+	5
	Intr	11990	12202	+	5
	Intr	12474	13163	+	5
	Intr	13195	14661	+	5
	Intr	14784	15536	+	5
20	Intr	15832	15986	+	5
	Intr	16190	16590	+	5
	Term	16653	17015	+	5
	Init	24295	24516	+	6
25	Intr	24596	24789	+	6
	Term	24862	25270	+	6
	Term	26197	26099	–	7
	Intr	26404	26306	–	7
30	Intr	27329	26629	–	7
	Intr	27606	27513	–	7
	Intr	27926	27760	–	7
	Intr	28093	28000	–	7
	Intr	28281	28172	–	7
35	Init	28752	28416	–	7
	Init	32354	34619	+	8
	Term	34765	35285	+	8
40	Term	37093	36987	–	9
	Intr	37340	37184	–	9
	Init	38662	38342	–	9
	Init	39578	39623	+	10
45	Intr	39702	41048	+	10
	Term	41181	41218	+	10
	Term	42494	41993	–	11
	Intr	43351	43047	–	11
50	Intr	44342	43449	–	11
	Intr	44765	44474	–	11
	Intr	46025	45145	–	11
	Intr	46704	46477	–	11
	Init	47832	47425	–	11
55					
	Init	49290	49457	+	12
	Intr	49648	49689	+	12
	Intr	49755	49946	+	12
	Intr	50218	50313	+	12
60	Intr	50394	50508	+	12
	Intr	51065	51070	+	12

858

Intr	51105	51192		+	12
Intr	51645	51853		+	12
Intr	51925	52040		+	12
Intr	52158	52307		+	12
Intr	52461	52528		+	12
Intr	52673	52778		+	12
Intr	52909	53067		+	12
Intr	53176	53229		+	12
Intr	53333	53413		+	12
Intr	53516	53653		+	12
Intr	53975	54134		+	12
Intr	54364	54443		+	12
Intr	54540	54688		+	12
Intr	54781	54838		+	12
Intr	54931	54978		+	12
Term	55141	55284		+	12
Init	55723	55846		+	13
Intr	55967	56162		+	13
Intr	56249	56476		+	13
Intr	56592	56823		+	13
Intr	56903	57010		+	13
Intr	57095	57367		+	13
Term	57445	57726		+	13
Init	60489	60741		+	14
Term	60774	60778		+	14
Term	62781	62035		−	15
Init	63693	62983		−	15
Sngl	66901	66713		−	16

>5541662   /

len =   98462   nex =   118

Init	108	225		+	0
Intr	324	434		+	0
Intr	606	659		+	0
Intr	824	1081		+	0
Intr	1184	1293		+	0
Intr	1483	1636		+	0
Intr	1707	2160		+	0
Intr	2248	2462		+	0
Intr	2558	2796		+	0
Intr	2889	3015		+	0
Term	3098	3310		+	0
Sngl	4862	3510		−	1
Sngl	5447	5947		+	2
Term	6581	6282		−	3
Intr	7204	6755		−	3
Init	8154	7357		−	3
Init	9457	11794		+	4
Intr	11847	12079		+	4
Intr	12162	12233		+	4
Intr	12305	12457		+	4

859

	Intr	12558	12721	+	4
	Intr	12816	12963	+	4
	Intr	13308	13430	+	4
	Intr	13523	13610	+	4
5	Term	13688	13884	+	4
	Term	14473	14138	–	5
	Intr	14673	14547	–	5
	Intr	14902	14708	–	5
10	Intr	15349	15012	–	5
	Intr	15542	15430	–	5
	Init	15678	15627	–	5
	Init	19089	19105	+	6
15	Intr	19139	19364	+	6
	Term	20408	20413	+	6
	Term	25081	24951	–	7
	Intr	25313	25191	–	7
20	Init	25427	25319	–	7
	Sngl	27284	27898	+	8
	Term	28131	28115	–	9
25	Intr	28300	28228	–	9
	Intr	28514	28414	–	9
	Intr	28813	28771	–	9
	Init	29021	28905	–	9
30	Term	30130	30075	–	10
	Intr	30681	30303	–	10
	Init	31303	30920	–	10
	Term	33911	33715	–	11
35	Init	34080	34020	–	11
	Term	35404	34872	–	12
	Intr	35734	35527	–	12
	Intr	35934	35823	–	12
40	Intr	36236	36021	–	12
	Intr	36412	36318	–	12
	Intr	37017	36507	–	12
	Intr	37361	37109	–	12
	Intr	37574	37483	–	12
45	Intr	37778	37684	–	12
	Init	37976	37881	–	12
	Term	39449	39084	–	13
	Intr	40739	40237	–	13
50	Intr	41067	40860	–	13
	Intr	41247	41136	–	13
	Intr	41543	41328	–	13
	Intr	41719	41625	–	13
	Intr	42328	41818	–	13
55	Intr	42677	42425	–	13
	Intr	42901	42810	–	13
	Intr	43101	43007	–	13
	Init	43303	43205	–	13
60	Init	43784	43787	+	14
	Intr	44307	44716	+	14

860

Intr	45919	46116	+	14
Term	47127	47615	+	14
Init	50322	50526	+	15
Intr	51049	51608	+	15
Term	51694	51867	+	15
Sngl	53782	55842	+	16
Init	57772	59044	+	17
Intr	59258	59566	+	17
Intr	60415	62335	+	17
Term	62797	62974	+	17
Term	63812	63293	−	18
Intr	64040	63982	−	18
Intr	64260	64192	−	18
Intr	64854	64387	−	18
Intr	65304	65132	−	18
Intr	65500	65365	−	18
Intr	66102	66047	−	18
Intr	69249	69084	−	18
Intr	69444	69332	−	18
Intr	69894	69537	−	18
Init	70250	70143	−	18
Term	72277	71274	−	19
Intr	73188	72978	−	19
Intr	73755	73262	−	19
Intr	74159	73845	−	19
Init	74370	74268	−	19
Init	75018	76208	+	20
Intr	76487	76708	+	20
Intr	76797	76904	+	20
Intr	77024	77215	+	20
Intr	77294	77542	+	20
Intr	77629	77731	+	20
Intr	77756	77883	+	20
Term	77971	78210	+	20
Sngl	78702	81245	+	21
Sngl	81966	83384	+	22
Sngl	85282	87825	+	23
Init	90153	90175	+	24
Term	90657	90984	+	24
Init	92466	92686	+	25
Intr	92718	92932	+	25
Term	93008	93885	+	25
Init	94553	94571	+	26
Intr	95083	95598	+	26
Term	95678	96555	+	26
Init	97572	98055	+	27
Term	98132	98441	+	27

>5541692   /

len =   87240   nex =   116

5	Sngl	987	56	−	0
	Init	2550	2892	+	1
	Intr	3699	4246	+	1
	Intr	4340	4714	+	1
10	Intr	4823	4858	+	1
	Intr	4958	5080	+	1
	Intr	5164	5271	+	1
	Intr	5362	5418	+	1
	Intr	5525	5587	+	1
15	Intr	5670	5911	+	1
	Intr	5989	6070	+	1
	Intr	6550	6721	+	1
	Intr	6860	6987	+	1
	Intr	7090	7232	+	1
20	Term	7436	7466	+	1
	Init	8989	8991	+	2
	Intr	9054	9847	+	2
	Intr	9905	9959	+	2
25	Intr	10165	10608	+	2
	Intr	10866	10972	+	2
	Intr	11253	11361	+	2
	Intr	11619	11738	+	2
	Intr	12009	12248	+	2
30	Intr	12329	12475	+	2
	Intr	12556	12646	+	2
	Intr	12725	12843	+	2
	Intr	12917	13087	+	2
	Term	13187	13234	+	2
35					
	Term	14074	13828	−	3
	Intr	15512	14193	−	3
	Intr	15766	15546	−	3
	Intr	15941	15891	−	3
40	Intr	16099	16049	−	3
	Init	16477	16181	−	3
	Init	20027	20332	+	4
	Intr	20850	21040	+	4
45	Intr	21121	21256	+	4
	Intr	21353	21669	+	4
	Intr	21770	21887	+	4
	Intr	22066	22215	+	4
	Intr	22282	22365	+	4
50	Intr	22457	22700	+	4
	Term	23295	23470	+	4
	Term	28207	28106	−	5
	Intr	28614	28296	−	5
55	Intr	28761	28697	−	5
	Init	28976	28833	−	5
	Sngl	29573	31780	+	6
60	Init	32880	33059	+	7
	Term	33139	33387	+	7

862

	Term	34092	33914	–	8
	Intr	34702	34168	–	8
	Intr	34931	34773	–	8
5	Intr	35815	35012	–	8
	Init	36333	36016	–	8
	Term	37787	37557	–	9
	Intr	37988	37881	–	9
10	Intr	38247	38080	–	9
	Intr	38465	38331	–	9
	Intr	38848	38546	–	9
	Intr	39146	38985	–	9
	Intr	39557	39225	–	9
15	Intr	39735	39640	–	9
	Intr	39962	39824	–	9
	Intr	40934	40854	–	9
	Init	43511	43450	–	9
20	Init	46269	46448	+	10
	Intr	46816	47008	+	10
	Intr	47119	47204	+	10
	Intr	47666	47797	+	10
	Intr	48085	48213	+	10
25	Term	48309	48515	+	10
	Init	50533	50715	+	11
	Intr	51061	51253	+	11
	Intr	51380	51465	+	11
30	Intr	51617	51748	+	11
	Intr	51837	51965	+	11
	Term	52058	52264	+	11
	Term	53234	53092	–	12
35	Intr	53480	53367	–	12
	Init	53858	53747	–	12
	Sngl	57216	57824	+	13
40	Init	61515	61871	+	14
	Intr	61974	62108	+	14
	Intr	62185	62284	+	14
	Intr	62360	62485	+	14
	Intr	62576	62889	+	14
45	Intr	62996	63115	+	14
	Intr	63221	63472	+	14
	Intr	63558	63683	+	14
	Term	63816	64076	+	14
50	Init	65050	65800	+	15
	Intr	66188	66523	+	15
	Term	66612	67309	+	15
	Sngl	68881	68216	–	16
55					
	Term	72539	71962	–	17
	Init	73600	72637	–	17
	Sngl	77944	78042	+	18
60					
	Init	80219	80406	+	19

863

	Type				
	Intr	80618	81087	+	19
	Intr	81693	81943	+	19
	Term	82028	82408	+	19
5	Term	82865	82616	−	20
	Intr	83056	82947	−	20
	Intr	83425	83324	−	20
	Intr	83625	83488	−	20
	Intr	83832	83714	−	20
10	Intr	84040	83995	−	20
	Intr	84363	84343	−	20
	Intr	84592	84511	−	20
	Intr	84768	84689	−	20
	Intr	84928	84842	−	20
15	Intr	85190	85005	−	20
	Init	85670	85272	−	20
	Init	86186	86503	+	21
	Term	86592	87207	+	21
20					

>5541715   /

len =   90780   nex =   145

	Type				
25	Term	72	1	−	0
	Init	462	159	−	0
	Term	1132	1087	−	1
	Intr	1388	1218	−	1
30	Intr	1587	1475	−	1
	Intr	2163	2041	−	1
	Intr	2356	2268	−	1
	Intr	2754	2622	−	1
	Init	3149	3063	−	1
35					
	Term	3745	3703	−	2
	Intr	3963	3841	−	2
	Intr	4336	4212	−	2
	Intr	4959	4837	−	2
40	Intr	5150	5062	−	2
	Intr	5388	5256	−	2
	Init	5644	5582	−	2
	Init	10758	11176	+	3
45	Intr	11291	11316	+	3
	Intr	11668	11698	+	3
	Intr	11844	11931	+	3
	Intr	12065	12213	+	3
	Intr	13021	13100	+	3
50	Term	13315	13562	+	3
	Term	14097	13924	−	4
	Init	16987	14453	−	4
55	Init	17210	17291	+	5
	Intr	17368	17833	+	5
	Intr	17968	18085	+	5
	Intr	18232	18341	+	5
	Intr	18427	18512	+	5
60	Intr	18635	18780	+	5
	Term	18898	18951	+	5

	Init	20458	21032	+	6
	Intr	21307	21436	+	6
	Intr	21812	21977	+	6
5	Intr	22063	22181	+	6
	Intr	22269	22349	+	6
	Intr	22527	22613	+	6
	Intr	22732	22806	+	6
	Intr	22896	22964	+	6
10	Intr	23038	23118	+	6
	Intr	23212	23346	+	6
	Intr	23436	23492	+	6
	Term	23602	23766	+	6
15	Init	26506	27657	+	7
	Term	27890	28204	+	7
	Term	31427	31153	−	8
	Intr	31771	31529	−	8
20	Intr	31980	31857	−	8
	Init	32877	32413	−	8
	Init	34620	35807	+	9
	Term	35890	36195	+	9
25					
	Term	38525	38336	−	10
	Intr	39060	38860	−	10
	Intr	40278	40070	−	10
	Intr	40543	40445	−	10
30	Intr	41749	40849	−	10
	Init	41968	41832	−	10
	Init	42452	42504	+	11
	Intr	42971	43715	+	11
35	Intr	43778	43918	+	11
	Term	44046	44312	+	11
	Sngl	44908	45429	+	12
40	Init	46357	46547	+	13
	Intr	46861	46963	+	13
	Intr	47051	47339	+	13
	Term	47431	47573	+	13
45	Init	49493	49622	+	14
	Intr	49792	49879	+	14
	Intr	49985	50051	+	14
	Term	50148	50201	+	14
50	Term	50889	50756	−	15
	Intr	51149	51021	−	15
	Intr	51385	51242	−	15
	Intr	51658	51533	−	15
	Intr	52044	51919	−	15
55	Intr	52226	52134	−	15
	Intr	52464	52329	−	15
	Intr	52672	52544	−	15
	Intr	52906	52787	−	15
	Intr	53227	53048	−	15
60	Intr	53500	53321	−	15
	Intr	53845	53711	−	15

865

Intr	54165	53953	–	15
Intr	54417	54329	–	15
Intr	54703	54554	–	15
Intr	54927	54789	–	15
Intr	55248	55027	–	15
Intr	55557	55372	–	15
Intr	55769	55647	–	15
Intr	55965	55865	–	15
Init	56473	56173	–	15
Init	57438	57490	+	16
Intr	57576	57615	+	16
Intr	57896	57976	+	16
Intr	58059	58253	+	16
Intr	58358	58468	+	16
Intr	58579	58644	+	16
Intr	58716	58823	+	16
Intr	58928	58972	+	16
Intr	59263	59319	+	16
Intr	59710	59804	+	16
Intr	60071	60257	+	16
Intr	60467	60529	+	16
Intr	60638	60744	+	16
Intr	60915	61052	+	16
Intr	61131	61212	+	16
Intr	61322	61392	+	16
Intr	61951	62031	+	16
Intr	62396	62453	+	16
Intr	62557	62609	+	16
Intr	62718	62811	+	16
Term	62904	62993	+	16
Term	64537	63431	–	17
Intr	65022	64609	–	17
Intr	65227	65111	–	17
Init	65755	65474	–	17
Term	66538	66497	–	18
Intr	66875	66789	–	18
Intr	67069	66980	–	18
Intr	67488	67383	–	18
Intr	67679	67585	–	18
Intr	67905	67822	–	18
Intr	68423	68344	–	18
Intr	68730	68690	–	18
Init	68954	68944	–	18
Sngl	69877	69737	–	19
Init	72249	72436	+	20
Intr	72522	72808	+	20
Intr	73121	73335	+	20
Intr	73435	73458	+	20
Intr	73818	73998	+	20
Intr	74078	74297	+	20
Intr	74390	74525	+	20
Intr	74768	74949	+	20
Intr	75101	75207	+	20
Term	75296	75378	+	20
Init	77483	79961	+	21

866

	Term	80157	81979	+	21
	Term	83660	83541	–	22
	Intr	83890	83750	–	22
5	Intr	84741	84446	–	22
	Init	85162	84859	–	22
	Term	89783	89043	–	23
	Init	90268	90095	–	23
10					

>5541718   /

len = 103889   nex =   195

15	Init	252	315	+	0
	Intr	471	526	+	0
	Intr	815	886	+	0
	Intr	989	1071	+	0
	Intr	1185	1260	+	0
20	Intr	1526	1615	+	0
	Intr	1717	1803	+	0
	Intr	1942	2013	+	0
	Intr	2125	2183	+	0
	Intr	2354	2447	+	0
25	Intr	2661	2879	+	0
	Intr	3037	3131	+	0
	Intr	3319	3409	+	0
	Intr	3511	3615	+	0
	Intr	3701	3739	+	0
30	Intr	3840	3929	+	0
	Intr	4080	4181	+	0
	Intr	4230	4502	+	0
	Term	4579	4677	+	0
35	Init	5850	6413	+	1
	Intr	6789	6840	+	1
	Intr	7556	7642	+	1
	Intr	7763	7827	+	1
	Intr	7931	8098	+	1
40	Term	8213	8263	+	1
	Term	8850	8813	–	2
	Intr	9269	9207	–	2
	Init	9758	9572	–	2
45					
	Init	12184	12210	+	3
	Intr	12368	12456	+	3
	Term	12862	13012	+	3
50	Term	13602	13513	–	4
	Intr	14969	13719	–	4
	Intr	15729	15417	–	4
	Init	16278	16193	–	4
55	Init	17354	17487	+	5
	Intr	17547	17632	+	5
	Intr	17722	17774	+	5
	Intr	17859	17936	+	5
	Intr	18047	18100	+	5
60	Intr	18192	18303	+	5
	Intr	18466	18533	+	5

867

Intr	18625	18700	+	5
Intr	18808	18869	+	5
Intr	18967	19372	+	5
Term	19459	19559	+	5
Term	20093	20003	−	6
Intr	20349	20243	−	6
Intr	20509	20449	−	6
Intr	20748	20597	−	6
Intr	20913	20836	−	6
Intr	21158	21045	−	6
Intr	21678	21602	−	6
Intr	21837	21771	−	6
Intr	22021	21956	−	6
Intr	22258	22135	−	6
Intr	22467	22346	−	6
Intr	22771	22508	−	6
Intr	22931	22851	−	6
Intr	23167	23025	−	6
Init	23631	23523	−	6
Sngl	24517	24284	−	7
Init	26696	26852	+	8
Term	26928	27058	+	8
Init	31341	33848	+	9
Intr	35122	35202	+	9
Intr	35576	35684	+	9
Intr	35771	35914	+	9
Intr	36259	36420	+	9
Intr	36531	36726	+	9
Intr	36817	36973	+	9
Intr	37093	37229	+	9
Intr	37326	37413	+	9
Intr	37768	37871	+	9
Intr	38193	38291	+	9
Intr	38940	38996	+	9
Intr	39256	39299	+	9
Intr	39839	39913	+	9
Intr	40514	40621	+	9
Intr	40805	40902	+	9
Intr	41034	41105	+	9
Intr	41529	41666	+	9
Intr	41765	41851	+	9
Intr	41932	42116	+	9
Intr	42485	43067	+	9
Intr	43150	43221	+	9
Intr	43353	43523	+	9
Intr	43567	43590	+	9
Intr	43719	43727	+	9
Intr	44199	44251	+	9
Intr	44887	45043	+	9
Intr	45175	45193	+	9
Intr	45233	45336	+	9
Intr	45497	45571	+	9
Intr	46046	46165	+	9
Intr	46372	46557	+	9
Intr	46768	46889	+	9
Intr	46968	47037	+	9
Intr	47132	47259	+	9

868

	Intr	47416	47557	+	9
	Intr	47663	47743	+	9
	Intr	47900	47998	+	9
	Intr	48089	48135	+	9
5	Intr	48656	48842	+	9
	Intr	48919	49140	+	9
	Intr	49641	49712	+	9
	Intr	50041	50110	+	9
	Intr	50608	50725	+	9
10	Intr	50822	50945	+	9
	Intr	51350	51484	+	9
	Intr	51726	51818	+	9
	Intr	52294	52341	+	9
	Intr	52429	52533	+	9
15	Intr	52794	52880	+	9
	Intr	53197	53514	+	9
	Intr	53801	53913	+	9
	Intr	54265	54331	+	9
	Intr	54420	54494	+	9
20	Intr	54993	55115	+	9
	Intr	55358	55522	+	9
	Intr	55600	55698	+	9
	Intr	56040	56191	+	9
	Intr	56312	56516	+	9
25	Intr	56589	56686	+	9
	Intr	56746	56797	+	9
	Intr	56921	57004	+	9
	Intr	57121	57204	+	9
	Intr	57944	58024	+	9
30	Intr	58141	58230	+	9
	Intr	58307	58369	+	9
	Intr	58648	58788	+	9
	Intr	58873	58929	+	9
	Intr	59017	59076	+	9
35	Intr	59536	59630	+	9
	Intr	59741	59849	+	9
	Intr	60042	60131	+	9
	Intr	60234	60313	+	9
	Intr	60412	60542	+	9
40	Intr	60589	60635	+	9
	Intr	60951	61082	+	9
	Intr	61204	61338	+	9
	Intr	61426	61491	+	9
	Intr	61733	61888	+	9
45	Term	61997	62074	+	9
	Term	62673	62419	−	10
	Intr	62994	62789	−	10
	Intr	63434	63209	−	10
50	Intr	64559	63570	−	10
	Init	66101	64899	−	10
	Term	67869	66662	−	11
	Intr	68153	67939	−	11
55	Intr	68343	68288	−	11
	Intr	68711	68421	−	11
	Intr	68988	68819	−	11
	Intr	69226	69084	−	11
	Intr	69465	69321	−	11
60	Intr	69658	69547	−	11
	Intr	70037	69760	−	11

869

```
        Intr   70277    70076          -        11
        Intr   70508    70372          -        11
        Intr   70740    70605          -        11
        Intr   71016    70872          -        11
5       Init   71960    71905          -        11

        Init   72785    72856          +        12
        Intr   73002    73120          +        12
        Intr   73173    73270          +        12
10      Intr   74410    74514          +        12
        Intr   76259    76444          +        12
        Intr   76771    76823          +        12
        Intr   77826    78297          +        12
        Term   78504    78616          +        12
15
        Init   79786    79994          +        13
        Term   80110    80146          +        13

        Init   83196    83331          +        14
20      Intr   83500    83565          +        14
        Intr   83712    83858          +        14
        Intr   84026    84139          +        14
        Intr   84233    84322          +        14
        Intr   84431    84553          +        14
25      Term   84672    85024          +        14

        Term   85977    85583          -        15
        Intr   86597    86026          -        15
        Init   87270    87227          -        15
30
        Init   89446    89496          +        16
        Intr   89677    89763          +        16
        Intr   89919    90131          +        16
        Intr   90240    90413          +        16
35      Intr   90537    90590          +        16
        Intr   91134    91280          +        16
        Term   91355    91732          +        16

        Init   98284    98541          +        17
40      Intr   98895    99184          +        17
        Term   99570    99972          +        17

        Init  101094   101250          +        18
        Term  101335   101720          +        18
45
```